University of Alberta

Computational Issues in Protein Nuclear Magnetic Resonance Spectroscopy

by

Xin Tu

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

Department of Computing Science

Edmonton, Alberta
Spring 2005

0-494-08187-2

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file *Votre référence*
ISBN:
Our file *Notre reference*
ISBN:

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Nuclear Magnetic Resonance (NMR) Spectroscopy is one of a few current techniques which can be used to determine three-dimensional structures of biomolecules at atomic resolution. The operating principle of NMR is based on the nuclear magnetic resonance phenomenon of the atomic nucleus. While it has been widely used to determine conformation of biomolecules, NMR experimentation is extremely time-consuming for large molecules such as proteins, due to their structural complexity. Due to its popularity, innovative data analysis and computational methods have been applied in NMR Spectroscopy. The time needed to resolve protein structure using NMR can be dramatically decreased from years to weeks with the assistance of these innovative methods. This thesis provides a comprehensive survey of NMR Spectroscopy including the operating principle of nuclear magnetic resonance, the detailed procedure for protein structure determination by NMR Spectroscopy, a comprehensive review of existing methods used in NMR Spectroscopy, and the advantages of applying computational methods.

Acknowledgements

I would like to thank:

My parents, for their support to me always;
My supervisor Guohui Lin, for his wise instructions and constant help;
My friends, for their encouragement.

Table of Contents

1 Introduction 1
- 1.1 NMR Spectroscopy . . . 1
  - 1.1.1 The Nuclear Magnetic Resonance Phenomenon . . . 2
  - 1.1.2 Chemical Shift . . . 5
- 1.2 Related Work . . . 6
  - 1.2.1 X-Ray Crystallography . . . 6
  - 1.2.2 Cryoelectron Microscopy . . . 8
- 1.3 Motivation . . . 9
  - 1.3.1 Protein Engineering . . . 9
  - 1.3.2 Drug Design . . . 10

2 Protein Structure Determination Procedure 11
- 2.1 NMR Instrumentation . . . 11
- 2.2 Data Acquisition . . . 14
- 2.3 Data Processing . . . 16
  - 2.3.1 Zero Filling . . . 16
  - 2.3.2 Apodization . . . 17
  - 2.3.3 Linear Prediction . . . 18
  - 2.3.4 Fourier Transform . . . 18
  - 2.3.5 Phase Correlation . . . 19
- 2.4 Peak Picking . . . 20
- 2.5 Peak Assignment . . . 22
  - 2.5.1 Peak Grouping and Adjacency Determination . . . 23
  - 2.5.2 Assignment Starting with Spin Systems . . . 24
  - 2.5.3 Assignment Starting with Spin Systems and Adjacency Constraint . . . 25
  - 2.5.4 Assignment Starting with Peak Lists . . . 27
  - 2.5.5 Scoring Scheme . . . 28
- 2.6 Structure Determination . . . 29
  - 2.6.1 Structural Constraints Extraction . . . 30
  - 2.6.2 Secondary Structure Determination . . . 34
  - 2.6.3 3D Structure Calculation . . . 38

3 Experimental Results **53**
3.1 Data Processing 53
3.2 Peak Picking 54
3.3 Peak Assignment 57
3.4 NOESY Assignment 58
3.5 3D Structure Calculation 60

4 Conclusions and Future Work **63**
4.1 Conclusions 63
4.2 Future Work 64

A Peak List Sample **71**

B NOE Distance Constraints File Sample **72**

C Dihedral Angle Constraints File Sample **73**

List of Figures

1.1 Possible orientations of nuclear precession axis 3
1.2 Alignment of nuclei under the external magnetic field 4
1.3 The energy difference between spin states 5
1.4 One unit cell in the crystal lattice 7
1.5 A crystal lattice is a three-dimensional stack of unit cells . . . 8

2.1 Schematic layout of an NMR spectrometer 12
2.2 Experimental scheme for quadrature detection 15
2.3 Nuclear overhauser effect patterns for various secondary structures 32
2.4 Hydrogen bond restraints 32
2.5 TALOS flowchart [18] 37
2.6 Variable target function algorithm 41
2.7 Tree structure of torsion angle dynamics algorithm DYANA . 46

3.1 One-dimensional Time-domain Spectrum FID 54
3.2 One-dimensional Frequency-domain Spectrum 54
3.3 Two-dimensional HSQC (heteronuclear single quantum correlation) Spectrum before Peak Picking 55
3.4 Two-dimensional HSQC (heteronuclear single quantum correlation) Spectrum after Peak Picking 57
3.5 Several conformers of protein MT0776 62

List of Tables

2.1 Distance bounds for different NOE intensity classes 31

2.2 Karplus relations, $^3J(\theta) = A\cos^2\theta + B\cos\theta + C$, for proteins between a vicinal scalar coupling constant ^3J and the corresponding torsion angle θ, defined by the three covalent bonds between the two scalar coupled atoms. "Offset" in the table represents the difference between θ and the standard torsion angle ϕ, ψ or χ^1. In the case of β-methylene protons, the first number is for $H^{\beta 2}$, the second for $H^{\beta 3}$. 34

2.3 CSI entries for the H^{α} chemical shifts [48] 36

2.4 Residue similarity factors, $\Delta_{ResType}$ [18] 38

2.5 Comparison of molecular dynamics simulations in Cartesian and torsion angle spaces 45

2.6 Computation time (in seconds) for DYANA structure calculations of the proteins BPTI and cyclophilin A on different computers [25, 27]. 52

3.1 Amino acid chemical shift statistics I [1] 59

3.2 Amino acid chemical shift statistics II [5] 59

Chapter 1

Introduction

The intention of this dissertation is to provide a comprehensive survey on Nuclear Magnetic Resonance Spectroscopy with a concentration on computational issues. We explain the Nuclear Magnetic Resonance phenomenon, and introduce the related applications of NMR spectroscopy. The focus of this dissertation is the complete procedure of applying NMR spectroscopy for resolving protein three-dimensional structures. In addition, we investigate a series of software to demonstrate the procedure of generating protein structures from NMR experimental data.

Chapter 1 describes the background knowledge and the motivation for this work, followed by a survey on some related work in the Nuclear Magnetic Resonance Spectroscopy field. Chapter 2 explains the procedures of NMR spectroscopy for protein structure determination. Chapter 3 demonstrates the complete process using a combination of different software, and presents the experimental results. Chapter 4 presents the conclusions and future work.

1.1 NMR Spectroscopy

NMR spectroscopy and X-ray crystallography are currently two major methods for protein structure determination. As an indirect experimental method, NMR spectroscopy might not be able to achieve the same accuracy as X-ray crystallography. The resolution of protein structure by X-ray crystallogra-

phy could reach less than 1Å, but 1-4Åfor NMR spectroscopy. Nevertheless, NMR spectroscopy is able to determine the three-dimensional structure of a protein in solution under nearly physiological conditions. Furthermore, the NMR spectroscopy process could be accelerated by computers and molecular modelling procedures, which have a larger development space. We believe that NMR spectroscopy will replace X-ray crystallography to become the dominant method of protein structure determination in the future. NMR spectroscopy was developed on the basis of the nuclear magnetic resonance phenomenon. In order to have a clear understanding of NMR spectroscopy, we need to review the essential principles of Nuclear Magnetic Resonance.

1.1.1 The Nuclear Magnetic Resonance Phenomenon

NMR spectroscopy is a protein structure determination technique in atomic resolution. The atom is the smallest unit in protein structure. Atoms consist of atomic nucleus and electrons, and the atomic nucleus consists of protons and neutrons. The atom has no charge, because the positive charge from each proton counteracts the negative charge from each electron, while neutrons have no charge. The *atomic number* is defined as the number of protons in the atom, which is used to distinguish different types of atoms. The same type of atoms with different numbers of neutrons are called *isotopes*.

A nucleus behaves as if it is spinning. The total effect of the imaginary spinning nucleus (protons and neutrons) is considered to be a physical property, the nuclear *spin*. Spin is an intrinsic angular momentum of the atomic nucleus which is measured by the nuclear spin quantum number I. Spins may be paired against each other in the nucleus. The rules to determine I are listed: (1) If either the neutron number or the proton number is odd, and the other one is even, the nucleus has a half-integer spin quantum number (i.e., 1/2, 3/2); (2) If both the neutron and the proton numbers are odd, the nucleus has an integer spin quantum number (i.e., 1, 2, 3); and (3) If the neutron and the proton numbers are both even, the nucleus has no spin, and the nuclei

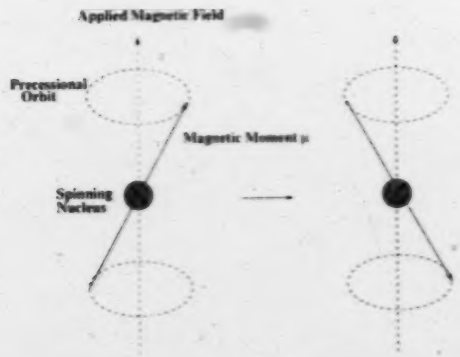

Figure 1.1: The precession axis of spinning nucleus changes from parallel (left) to opposite (right) orientation

in this category are NMR-inactive. For a nucleus of spin I, there are $2I + 1$ spin states (energy levels), ranging from $-I$ to $+I$ which corresponds to the magnetic spin *quantum* number m:

$$m = -I, -I + 1, ..., I - 1, I.$$

The nuclei with $I = 1/2$ include ^{1}H (Hydrogen), ^{13}C (Carbon), ^{15}N (Nitrogen), ^{19}F (Fluorine) and ^{31}P (Phosphorus); the nucleus with $I = 1$ includes deuteron ^{2}H; and the typical isotopes with no spin include ^{12}C, ^{14}N, ^{16}O (Oxygen). The nuclei with $I = 1/2$ are the easiest ones to detect in NMR spectroscopy experiments, so the following text is concerned only with the NMR spectroscopy of spin $I = 1/2$ nuclei which have two spin states: high and low energy.

Any charged object has a magnetic moment, and generates a magnetic field when it is moving. Therefore, the spinning nucleus produces a small electric current and a magnetic field associated with it. The magnetic moment μ of the spinning nucleus is a vector represented by the arrows, as seen in Figure 1.1. The natural state of nuclei in atoms is shown in the left side of Figure 1.2. When an external magnetic field is applied to the spinning nuclei, the nuclei in the field will precess about the spin rotation axis with the angular frequency ω_0, i.e. *Larmor Frequency*, as shown in Figure 1.1. The Larmor Frequency is given by

$$\omega_0 = \gamma B_0,$$

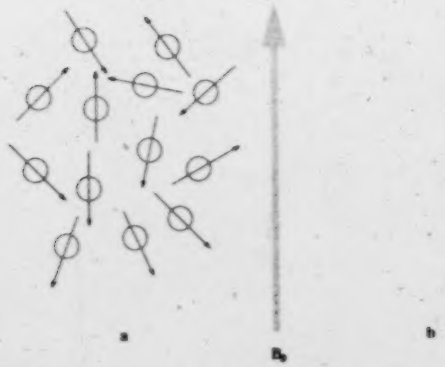

Figure 1.2: (a) Nuclei in the natural state; (b) Nuclei under the external magnetic field B_0

where the gyromagnetic ratio γ is a characteristic constant for a given nucleus, and the precession frequency ω_0 is proportional to the strength of the external magnetic field B_0. There will be only two possible orientations of the precession axis for the nucleus under the external magnetic field: parallel (low energy) or opposite (high energy) to the external field, as illustrated in Figure 1.1.

When the precessing nuclei are irradiated with a radio frequency (RF) pulse at the proper frequency, nuclei will spin-flip either from low energy state to high energy state or from high energy state to low energy state, by absorbing or emitting a certain quantum of energy (Figure 1.3). When this spin transition occurs, the nuclei are said to be in resonance with the applied radiation, which is referred to as *Nuclear Magnetic Resonance*. The electromagnetic radiation supply can induce nuclear magnetic resonance only if its frequency is equal to the frequency of the oscillating electric field generated by nucleus precession. Under these circumstances, the energy needed in resonance can be transferred from the electromagnetic radiation to the precession nucleus. As illustrated in Figure 1.3, the absorbing or emitting energy is equal to the energy difference between two spin states, which is expressed by

$$\Delta E = E_1 - E_2 = \hbar\omega_0 = \gamma\hbar B_0, \tag{1.1}$$

where $\hbar$ is Planck's constant h divided by 2π, and E_1 and E_2 denote the energy levels of nuclei in the spin states $-1/2$ and $1/2$, respectively. As the frequency

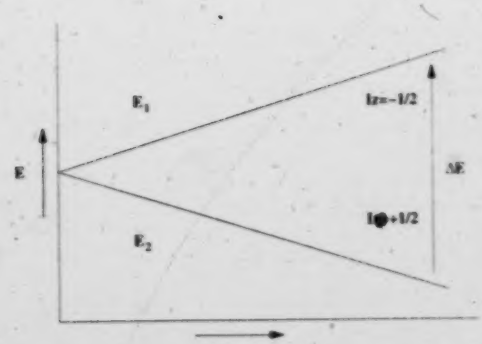

Figure 1.3: The energy difference between spin states, E_1 and E_2 represents high energy and low energy level

of the nuclear transition between two energy levels, v_0 can be calculated by

$$v_0 = \frac{\omega_0}{2\pi} = \frac{\gamma B_0}{2\pi}, \tag{1.2}$$

which is proportional to the angular Larmor frequency ω_0. Therefore, Equation (1.1) could also be written as $\Delta E = hv_0$. During the absorbing procedure, it is possible that the nuclei will reach a state with equal population of the nuclei in both directions; the spin system is then saturated. Therefore, the existence of the saturated state implies the presence of the relaxation process. Without further RF pulses, the spin system will return to the thermal equilibrium by the relaxation process. During the relaxation process, the decaying precession frequency is recorded and amplified by NMR instrumentations. The resulting signal is "Free Induction Decay" (FID). Further processing of FID will be discussed in later sections.

1.1.2 Chemical Shift

The resonance frequencies of each individual nuclei are not only relevant to the strength of the external magnetic field B_0 applied to it, but also dependent on the local chemical environments of the individual nuclei. The secondary magnetic field is generated by the motion of the electrons surrounding the nucleus, induced by the external magnetic field. The actual magnetic field depends on

both the external magnetic field and the secondary magnetic field. The modulation effect of the secondary field is defined as *shielding*. The variation of the resonance frequency with shielding is termed as *chemical shift*. The actual field present at the nucleus is denoted by B_{local} which is equal to $B_0(1-\sigma)$, and the term shielding is represented by the Greek letter σ. If B_0 is replaced by B_{local} in Equation (1.2), the resonance frequency is defined as

$$v_0 = \frac{\gamma B_0(1-\sigma)}{2\pi}.$$

Chemical shift in Parts Per Million (PPM) is defined as

$$\delta = \frac{(\omega_0 - \omega_{\text{reference}}) \times 10^6}{\omega_{\text{reference}}} \approx (\sigma_{\text{reference}} - \sigma) \times 10^6,$$

where ω_0 is the resonance frequency of the nucleus, and $\omega_{\text{reference}}$ is the reference frequency. For both proton and carbon, the reference frequency is the resonance frequency relative to that of the tetramethylsilane ($(CH_3)_4Si$, usually called TMS). Chemical shift makes the nuclei in different chemical environments distinguishable. Chemical environment refers to the interactions between nuclei, including chemical bonds, scalar coupling, dipolar coupling and hydrogen bond. On the basis of the chemical shift phenomenon, it is possible to map different chemical shifts to amino acid residues by peak assignment. This is one of the key processes of NMR spectroscopy for protein structure determination. The detailed procedure for applying Nuclear Magnetic Resonance in protein structure determination will be presented in the next chapter.

1.2 Related Work

1.2.1 X-Ray Crystallography

X-ray crystallography is an existing technique for protein structure determination used as an alternative to NMR spectroscopy. It is a technique developed on the basis of physical scattering properties of electrons. The basic principle

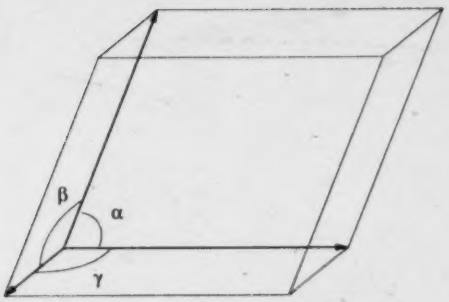

Figure 1.4: One unit cell in the crystal lattice

of X-ray crystallography is to place the crystal of a protein in X-rays, and the X-rays diffracted from the crystal can be recorded on a piece of photographic film with the angles and intensities information. The three-dimensional structures of proteins are determined by results of the analysis of reflections on the photographic films.

In X-ray crystallography, all the experiments are based on crystals of proteins. Therefore, the first step is to grow a crystal of the protein. The crystal can be described as a three-dimensional heap of unit cells with their edges forming a crystal lattice. A unit cell is the smallest repeating unit in the crystal, it is defined by three vectors — a, b, c — which denote three lattice dimensions. These three vectors (a, b, c) can be determined by angles between them which are denoted as α, β, and γ. To repeat a unit cell in a crystal lattice, α, β, and γ are necessary parameters. The crystal lattice structure is depicted in Figure 1.5.

The crystal is placed in the X-ray beam after it has been prepared. Under the radiation of X-ray, the crystal may be damaged by the energy passing from the X-ray. To avoid such damage, in the experiments, experts use a nitrogen stream "cryostream" to keep the crystal at very low temperatures. The source of information used in X-ray Crystallography is the diffraction pattern of X-rays being deflected by the crystal. In the experiments, it is necessary to rotate the crystal at such an angle to provide sufficient data on different angles. Diffraction patterns of X-rays include not only the intensity and the position of the spots, but also the phases of the deflected rays. The

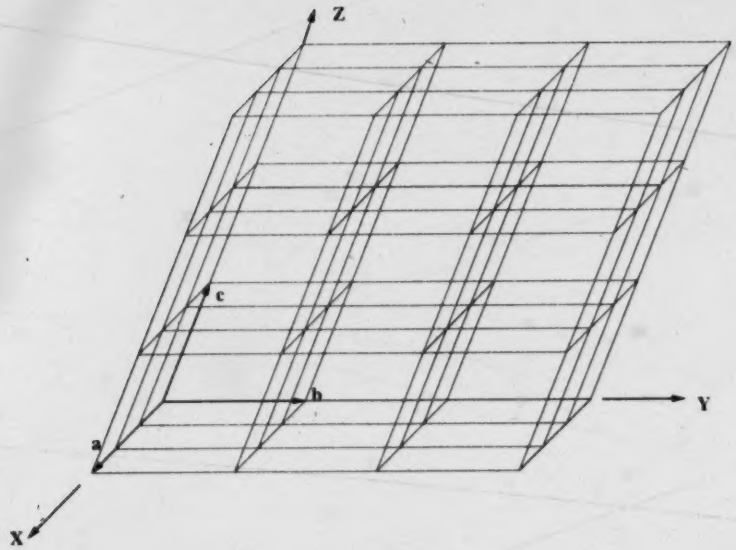

Figure 1.5: A crystal lattice is a three-dimensional stack of unit cells

relative positions of two atoms can be reflected from the phase shift between two deflected rays.

From a computing science point of view, if the phase information of the spots is known, then corresponding electron densities can be calculated. A three-dimensional electron density is generated to map to amino acid groups in the protein sequence.

1.2.2 Cryoelectron Microscopy

Although it is one of the protein structure determination techniques, cryoelectron microscopy is no longer widely used in this area. It is a three-dimensional image reconstruction technique that involves freezing the biological sample in order to view the sample with the least possible distortion and the fewest possible artifacts. Cryoelectron microscopy is particularly designed for proteins whose crystals are difficult to obtain, such as large multisubunit proteins. The structures of these proteins can be obtained by cryoelectron microscopy.

In this technique, a protein sample is rapidly frozen in liquid helium to

preserve its structure, and then examined in the frozen, hydrated state in a cryoelectron microscope. Micrographs are taken and recorded on film. A micrograph looks like a photograph of a region of the ice with the particles frozen in mid-tumble. Sophisticated computer programs analyze two-dimensional micrographs and reconstruct the protein's structure in three dimensions.

1.3 Motivation

As is well known, the functions of protein are determined by protein structures. Techniques for protein structure determination play an essential role in many fields related to protein functions. As one of the two main experimental methods for protein structure determination, NMR spectroscopy may not be able to achieve the same accuracy as X-ray crystallography; however, NMR spectroscopy complements it in many ways. With the application of computational methods and dynamics information with the protein functions, NMR spectroscopy becomes a more efficient technique for protein structure determination, and is expected to play a more significant role in structural biology. The motivation for NMR spectroscopy research comes from the rapid growth of applications of protein. The rest of this section will describe two essential applications of protein structure determination.

1.3.1 Protein Engineering

Protein engineering is based on protein's three-dimensional structure and its relationship to biological functions. Its purpose is to reconstruct the natural proteins and, further, create proteins more suitable for human needs through molecular design and its subsequence on gene transformation. In brief, the aim of protein engineering is to reconstruct natural proteins in order to create required proteins with specific functions for applications in industry, agriculture, and medical treatment, etc.

In protein engineering, the first step is to determine protein structures. Once the protein structure is ready, it is added to protein databases, such as PDB (Protein Data Bank) to compare with other proteins, in order to discover patterns of different levels of the protein structure which are relevant to its biological functions. From this step, we can discover the inherent relationship between the structures and functions of proteins.

There are several examples of research into protein engineering— such as insulin protein engineering and antibody humanization. Insulin is a specific medicine for diabetes, however, natural insulin lasts only a few hours after injection. Patients who require several injection a day experience more pain. It is important, therefore, to create a new type of insulin with more long-lasting and beneficial effects, in order to minimize patient discomfort. In insulin protein engineering, one way to achieve this is to keep the biologically active structure and reinforce the connection strength among other parts of the structure at the same time, in order to resist damage from the enzyme.

1.3.2 Drug Design

Another application of protein is the drug design based on macro-molecules, such as protein and nucleic acid. These macro-molecules are primarily proteins. Medicine takes effect through the interaction between the medicine and the receptor. Therefore, once the protein structure of the receptor is known, we can create the most effective medicine to transform its structure to produce a cure. There exist many kinds of drug designs based on macro-molecules. There are drug designs based on the structure of the antibody, the cancer gene expression product, and the structure of the receptor on the surface of the cell, for example. In principle, as long as structures of macro-molecules with biological functions are known, we can design the complementary small molecules to adjust their functions.

Chapter 2

Protein Structure Determination Procedure

The physical principles of Nuclear Magnetic Resonance (NMR) phenomena and related works of NMR spectroscopy were presented in Chapter 1. This chapter will introduce the experimental procedure for protein structure determination using NMR spectroscopy. Topics to be discussed include NMR instrumentation, data acquisition, data processing, peak picking, peak assignment, and structure determination.

2.1 NMR Instrumentation

A wide variety of NMR instrumentation is available for experiments. The common components of NMR spectrometers are: (a) superconducting magnet for supplying an external magnetic field B_0, (b) pulse programmer and radio frequency (RF) transmitter to generate and control RF pulse with strength B_1, (c) probe for placing the sample in the magnet, (d) receiver for receiving the resulting NMR signals, and (e) computers for data acquisition and processing. These subsystem components are shown in Figure 2.1.

The magnet component consists of a field-frequency lock system, shim coils, and probe. The field-frequency lock system built into the magnet is used to maintain the stability of the static magnetic field. The shim coil is utilized to optimize the spatial homogeneity of the magnetic field by integrating a set

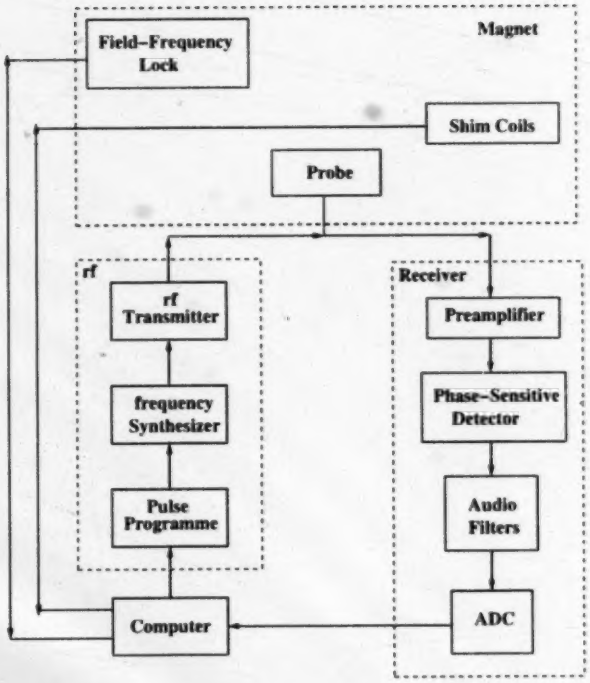

Figure 2.1: Schematic layout of an NMR spectrometer. The major components, including the magnet, RF electronics, receiver, and computer; and important subsystems are illustrated [14].

of auxiliary room-temperature electromagnets into the magnet component to compensate the inhomogeneity of the main magnetic field. This process is usually called "shimming". The probe is used to generate the main magnetic field B_0.

The RF transmitter is composed of a frequency synthesizer, amplifiers, and pulse programmer. Among these RF electronics, the frequency synthesizer is the source of the RF pulses. The pulse programmer here is used to control the timing, duration, amplitudes, and phases of the RF pulses. In NMR experiments, usually one RF transmitter is needed for generating proton frequencies. Additional RF transmitters are needed for generating RF frequencies for heteronuclear spectroscopy.

The receiver component in the figure contains the preamplifier, phase-sensitive detector, and analog-to-digital converter (ADC). The preamplifier is used to amplify the weak NMR signals from the probe. The phase-sensitive detector provides quadrature detection of the signal. Audio filters included in this component are used to limit the frequency bandwidth of the spectrometer. The ADCs convert the analog signals into digital signals for digital processing in the computer.

Early NMR experiments were conducted on electromagnets and operated in the continuous wave (CW) mode. NMR instruments using a superconducting magnet and operating in the pulse Fourier transform (FT) mode now dominate the market. All NMR spectroscopy experiments discussed in this thesis are based on this type of NMR instrument. The advantages of a superconducting magnet over an electromagnet are its higher sensitivity and greater stability. This is because the differences between the chemical shifts are amplified, with the increase of magnetic field strength, which leads to better separation between nuclei resonances.

2.2 Data Acquisition

In the spectrometer, a superconducting magnet provides static magnetic field B_0. Transverse magnetic field B_1 will be generated by a series of RF pulses. The time-varying current resulting from the probe is amplified and digitized by the preamplifier and ADC, respectively, and then recorded by the spectrometer. The resulting current and time signal are called free-induction decay (FID). FID will be sent to the computer for further processing to be introduced in the next section.

The time domain signal is sampled at evenly spaced time intervals, with the sampling interval denoted by Δt. The sampling rate is

$$\Delta t^{-1} = 2f_n.$$

Here, f_n represents the *Nyquist frequency*. To understand the *Nyquist frequency*, it is necessary to introduce the *Nyquist-Shannon sampling theorem*.

Theorem 2.2.1 (Nyquist-Shannon Sampling Theorem [4]) *When converting from an analog signal to digital (or otherwise sampling a signal at discrete intervals), the sampling frequency must be greater than or equal to twice the highest frequency of the input signal in order to be able to reconstruct the original perfectly from the sampled version.*

If the sampling rate is $2f_n$, all the components with frequencies higher than f_n will be aliased or folded to lower-frequency components. The *Nyquist frequency* f_n is the maximum frequency of the components which can be reconstructed with sampling rate $2f_n$.

To simplify the digitization process, the magnetic resonance frequencies in NMR spectroscopy are measured as their differences from a reference frequency. If the resonance frequency is larger than the reference frequency, the difference frequency is positive; otherwise, the difference frequency is negative. The problem with the digitization process is that this process only records the magnitude of frequency difference, but not the sign. It is impossible to distinguish +500 Hz or -500 Hz from the reference frequency. The problem can

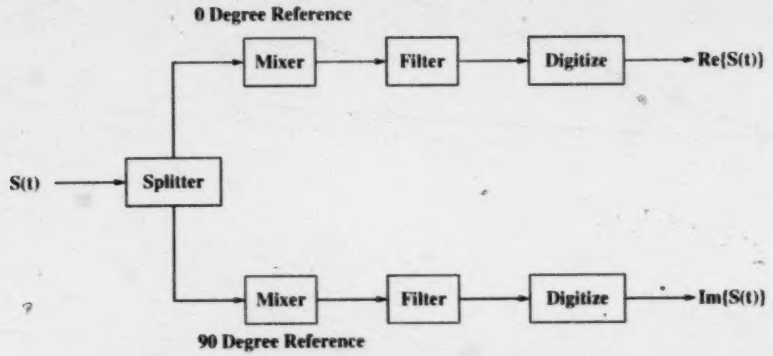

Figure 2.2: Experimental scheme for quadrature detection. The incoming signal recorded by the probe and preamplifier is split into two parallel channels. The signal in each channel is mixed with a reference signal, passed through a low-pass audiofilter, and digitized. Real (cosine-modulated) and imaginary (sine-modulated) components of the signal are obtained by shifting the relative phase of the reference signals by 90° [14].

be solved by placing the reference frequency to one side of the frequency spectrum, whereby all the resonance frequencies have the same sign. However, the disadvantages of placing the reference frequency to one side instead of in the middle of the spectrum include: (1) increased noise level of the spectrum; (2) requirement for more RF transmitter power due to the doubled frequency range generated from RF pulses; and (3) increase of the requirements in the data acquisition instrument.

As concluded from the above, the reference frequency will be kept in the middle of the frequency spectrum. A technique called "Quadrature Detection" is then introduced to determine the sign of the frequency. Quadrature detection records both sine and cosine components of the signals during sampling, as shown in Figure 2.2. The two channels are cosine-modulated and sine-modulated signals with frequency $\omega_0 - \omega_{ref}$. If the signal resulting from the probe and preamplifier is sine-modulated as $\cos(\omega_0 t)$, the quadrature detection process [14] can be described by:

$$\cos(\omega_0 t) \xrightarrow{splitter} \cos(\omega_0 t) - i\cos(\omega_0 t)$$

$$\xrightarrow{mixers} \cos(\omega_0 t)\cos(\omega_{ref} t) - i\cos(\omega_0 t)\sin(\omega_{ref} t)$$

$$
\begin{aligned}
&= && \frac{1}{2}\cos[(\omega_0+\omega_{ref})t] + \frac{1}{2}\cos[(\omega_0-\omega_{ref})t] \\
& && -\frac{1}{2}i\sin[(\omega_0+\omega_{ref})t] + \frac{1}{2}i\sin[(\omega_0-\omega_{ref})t] \\
& \xrightarrow{audiofilters} && \frac{1}{2}\cos[(\omega_0-\omega_{ref})t] + \frac{1}{2}i\sin[(\omega_0-\omega_{ref})t] \\
&= && \frac{1}{2}\exp[i(\omega_0-\omega_{ref})t],
\end{aligned}
$$

in which $i = \sqrt{-1}$ is a mathematical mechanism to distinguish the signals between two channels.

2.3 Data Processing

As illustrated in Figure 2.1, the digital signals resulting from ADC will be sent to the computer for further processing. The typical technique used in data processing is the Fourier Transform which extracts information from the digital signals. A variety of other data processing techniques are applied before or after Fourier Transform to optimize or purify the results.

Multiple processing methods applied on the NMR data preparation are: (1) zero filling, (2) apodization, (3) linear prediction, (4) Fourier transform, and (5) phase correlation. In the rest of this section, a detailed introduction to each processing method will be given.

2.3.1 Zero Filling

Zero filling is one of the data processing techniques used prior to Fourier transform. The reason for this step is that the complex Fourier transform of N data points consists of a real component and an imaginary component. Only half of the original number of data points are included in the real spectrum component, and information is lost while discarding the imaginary spectrum. The resulting spectrum has only half the resolution of the original one. Therefore, zero filling is designed to improve the resolution of the spectrum by appending

zeros after the collected data points. In addition, if the fast Fourier transformation (FFT) is used to calculate the discrete Fourier transformation, zero filling is needed to change the number of data points to be an integral power of 2, i.e., 2^n for some integer n.

If N is equal to 2^n for some integer n, zero filling will double the number of data points by filling zeros. If N is not an integral power of 2 but in the range $(2^n, 2^{n+1})$, zero filling will generate a new data set of 2^{n+1} data points by appending $(2^{n+1} - N)$ zeros to the end of the original data set. No information is obtained from additional zeros appended in the data set. Further improvement of the appearance of the spectrum can be achieved by interpolating zeros between data points in the frequency domain spectrum.

2.3.2 Apodization

Apodization is another technique used to optimize the spectrum before applying Fourier transform. Apodization is utilized to reduce truncation artifacts, increase signal-to-noise ratio, and generate desired lineshapes. Apodization works by multiplying the apodization function $A(\omega)$ to the original lineshape function in frequency domain $S(\omega)$ as Equation (2.1)

$$S'(\omega) = A(\omega) * S(\omega). \tag{2.1}$$

Equivalent function in time domain $a(t)$ will generate the same lineshape when it is multiplied with continuous function $s(t)$ in time domain as Equation (2.2)

$$S'(\omega) = \mathcal{F}\{a(t)s(t)\}. \tag{2.2}$$

Here, $\mathcal{F}$ represents the Fourier transform of the function $a(t)s(t)$. In order to reduce the truncation artifact, the time-domain spectrum FID has been brought smoothly down to zero by apodization at time t_{max}. Therefore, the corresponding frequency-domain lineshape has been broadened, and the resolution of the spectrum decreases.

There are a variety of apodization functions available, such as "Bartlett", "cosine", "Gaussian", "Hamming", "Hanning", "uniform", and "Welch"[2].

For instance, function "cosine" is calculated as

$$a(t)_{cosine} = \cos(\pi t/2t_{max}).$$

2.3.3 Linear Prediction

The most typical use of the linear prediction (LP) is to interpolate missing data points or to extrapolate the FID beyond the acquisition time, in order to improve spectra resolution and increase signal-to-noise ratio. The k-th data point can be predicted from the linear combination of the preceding M data points by

$$s(k\Delta t) = -\sum_{m=1}^{M} a_m s([k-m]\Delta t) + \varepsilon_m,$$

in which M is the number of previous data points, i.e., the prediction order of LP; k is an integer in the range $[0, N-1]$; a_m is the m-th linear prediction coefficient; and ε_m represents the prediction error. The optimal solution of LP is the set of the linear prediction coefficient a_m which minimizes the least-squares of the prediction errors, i.e.,

$$\min\left(\sum_{m=1}^{M} \varepsilon_m^2\right).$$

In general, LP is applied to the FID before Fourier transform in order to optimize the resulting frequency-domain spectrum.

Furthermore, linear prediction is used as the alternative of the Fourier transform. The frequency-domain spectrum can be generated directly from the set of linear prediction coefficients, a_m. Owing to the higher computational complexity of the linear prediction method compared to the Fourier transform, it is not feasible to use LP for the case of large data sets.

2.3.4 Fourier Transform

Fourier transform is responsible for transformation between the time-domain and the frequency-domain function which is an important step in data pro-

cessing. Typically, Fourier transform is to transform the time-domain function to the frequency-domain function as

$$\begin{aligned} S(\omega) &= \mathcal{F}\{s(t)\} = \int_{-\infty}^{\infty} s(t)e^{-i\omega t}dt; \\ S(v) &= \mathcal{F}\{s(t)\} = \int_{-\infty}^{\infty} s(t)e^{-i2\pi vt}dt, \end{aligned} \tag{2.3}$$

in that, $\omega = 2\pi v$. The inverse Fourier transform is the inverse process of the typical Fourier transform:

$$\begin{aligned} s(t) &= \mathcal{F}^{-1}\{S(\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)e^{i\omega t}dw; \\ s(t) &= \mathcal{F}^{-1}\{S(v)\} = \int_{-\infty}^{\infty} S(v)e^{i2\pi vt}dv. \end{aligned} \tag{2.4}$$

Since discrete Fourier transform has lower computational complexity than direct Fourier transform, and the FID has been digitized by sampling, discrete Fourier transform is applied in NMR experiments:

$$S(v) = S[k/(N\Delta t)] = \mathcal{F}\{s(j\Delta t)\} = \sum_{j=0}^{N-1} s(j\Delta t)e^{-i2\pi jk/N},$$

and the inverse process of the discrete Fourier transform is described as

$$s(j\Delta t) = \mathcal{F}^{-1}\{S[k/(N\Delta t)]\} = \frac{1}{N}\sum_{k=0}^{N-1} S[k/(N\Delta t)]e^{i2\pi jk/N}.$$

In the two equations above, N is the number of data points, Δt is the sampling interval, and $k = -N/2, ..., 0, ..., N/2$. The frequency range represented by the Fourier transformed spectrum is $[-f_n, f_n]$, in which Nyquist frequency f_n is equal to $1/(2\Delta t)$.

No information is lost during the Fourier transform because the signal energy is identical in both time-domain and frequency-domain signals, based on Parseval's theorem [2].

2.3.5 Phase Correlation

The real component of the frequency-domain spectrum consists of absorptive and dispersive lineshapes denoted by $A(\omega)$ and $D(\omega)$, respectively, and they

are calculated by

$$\begin{aligned} A(\omega) &= \cos[\theta_0 + \theta_1(\omega)]Re\{S(\omega)\} + \sin[\theta_0 + \theta_1(\omega)]Im\{S(\omega)\}; \\ D(\omega) &= -\sin[\theta_0 + \theta_1(\omega)]Re\{S(\omega)\} + \cos[\theta_0 + \theta_1(\omega)]Im\{S(\omega)\}, \end{aligned}$$

in which θ_0 is the zero-order phase correction and $\theta_1(\omega)$ is the first-order phase correction. The function of the phase correlation is to optimize the appearance of the frequency-domain spectrum by adjusting two parameters, θ_0 and $\theta_1(\omega)$, until the lineshapes in the real component of the spectrum are absorptive only.

2.4 Peak Picking

Peak picking is a process designed to filter out artificial peaks, to calibrate NMR signal lineshapes, and recognize the intensity of each peak. A series of two- and three-dimensional NMR spectra are used for protein structure determination, where each dimension represents the chemical shift in ppm for a certain type of nuclei. Because of strongly overlapping peaks, and spectral distortions due to artificial peaks, even robust methods might fail for NMR signal recognition in complex spectra.

There exists a variety of software for performing the peak picking process. For example, AUTOPSY [35] is able to deal with overlap and deviations from ideal Lorentzian line shape. ATNOS [31] was developed mainly for automated Nuclear Overhauser Effect Spectroscopy (NOESY) peak picking.

In more detail, AUTOPSY is an automated peak picking and peak integration method. The essences of this program are the function for local noise level calculation, the use of lineshapes extracted from well-separated peaks for resolving strongly overlapping peaks, and the consideration for symmetry. The key observation utilized by AUTOPSY is that multidimensional spectra typically contain multiple peaks that have the same lineshape and the same chemical shift in one frequency domain.

When using AUTOPSY, the first step is to determine noise level, which is crucial so that weak peaks can also be recognized. The noise level is calculated

locally since the noise level value for each 1D slice (rows and columns in 2D spectra) is calculated. Therefore, AUTOPSY is able to deal with water lines and artifacts. The second step is the segmentation process, which divides the spectrum into connected regions composed of data points; the data points out of the regions are discarded in this step. The well-separated peaks are identified first in the third step, using a symmetry violation criterion. In order to resolve the duplication problem of the lineshapes caused in the previous step, the fourth step is to group the similar lineshapes. Then, in the fifth step, the group of lineshapes are used to generate the potential peaks, which are a combination of the lineshapes from each dimension. This process is restricted in a bounding box of the region. The generated peaks with centers outside the bounding box are not processed any further. A list of potential peaks is compared with the recognized peaks; if there is a match, the peak in the list is marked as "found". The last step is the symmetrization check process. A quality factor is introduced as a criterion for further processing of the peaks. Only the peaks with a quality factor reaching the requirement are processed further. By using this procedure for peak picking, AUTOPSY is able to resolve most overlap, artifacts, and deviation of the lineshapes.

ATNOS is an automated NOESY peak picking software used to extract structural constraints. The input to ATNOS includes target protein sequence, chemical shift lists from peak assignment, and several 2D or 3D NOESY spectra. The current implementation of ATNOS performs multiple cycles of NOE peak identification, combined with automated NOE assignment program CANDID [32]. In each cycle, ATNOS performs automated NOE peak picking using NOESY symmetry criterion. By re-assessing the NOESY spectra in each cycle of structure calculation, ATNOS enables direct feedback among the NOESY spectra, the NOE assignments, and the protein structure. The new software package RADAR which combines ATNOS and CANDID will be freely available soon according to [32].

2.5 Peak Assignment

Through NMR data acquisition, processing, and spectral peak picking, the output from every NMR experiment is a list of spectrum peaks. In practice, the output peak list might still contain artificial peaks and some real peaks might be missing from the list because of data degeneracy, which may be corrected in the peak assignment. This list of peaks will be mapped to their host nuclei in the amino acid residues in the target protein sequence. Such a process is referred to as *NMR peak assignment* involving multiple peak lists. As one of the crucial processes, even a small error in peak assignment may cause a huge structure gap in the result. While NMR knowledge still plays an essential role in the success of peak assignment, computational techniques become increasingly important in the automation of the peak assignment process, ensuring that the peak assignment will become a high-throughput process in the near future.

Peak assignment is used to map peaks from multiple spectra to their host nuclei in the target protein sequence. The two main pieces of information used in this process are *signature information* and *adjacency information*. Signature information is the statistical information about nuclei, and the local chemical environment and chemical shift values. In other words, each type of nuclei in a specific local chemical environment will have chemical shift values in a very narrow range. Adjacency information comes from the correlation between multiple spectra. The chemical shift of the common nucleus in multiple spectra will be used to bridge peaks in those spectra, and primarily used to connect adjacent spin systems. A spin system refers to a set of chemical shifts which arise from nuclei residing in a common amino acid residue. Therefore, peak assignment is another phase for detecting artificial peaks if they are in conflict with the formed spin systems and their mapped residues.

There are many free software tools available for peak assignment. To enumerate a few: PASTA [37] uses threshold accepting algorithms; GARANT [8, 9] uses genetic algorithm; PACES [15] and MAPPER [29] use exhaustive

search algorithms; AutoAssign [53] uses heuristic best-first search algorithms. Some do the peak grouping and adjacency determination first, and then move on to the spin system assignment under the adjacency constraints; others do the peak grouping and adjacency determination, and the chained spin system assignment at the same time. Various peak assignment methods are introduced in the following sections.

2.5.1 Peak Grouping and Adjacency Determination

In the various assignment methods to be discussed, the peak grouping and adjacency determination processes are basically the same. First, we will describe how to group peaks into spin systems and the process of adjacency determination, using the spectra of small proteins as an example. Three spectra are used to demonstrate the process: 2D 15N-HSQC, 3D CBCAcoNH, and 3D HNCACB.

In 2D 15N-HSQC, each peak contains two entries: one for the amide proton (HN) chemical shift, the other for the directly attached nitrogen (NH) chemical shift. In 3D CBCAcoNH, each peak contains three entries — for HN, NH, and carbon chemical shift, respectively. The entry for carbon includes carbon alpha (CA) and carbon beta (CB) from the same amino acid residue. Those three entries are the same as in the 3D HNCACB; the difference is the carbon entry, which includes CA and CB from two adjacent amino acid residues in 3D HNCACB. HN and NH chemical shift values in the spin systems are from 2D 15N-HSQC. CA and CB chemical shift values are then extracted from CBCAcoNH and HNCACB which are complementary to each other. However, in general, the chemical shifts measured out of one NMR spectrum are different from those measured out of another. Nonetheless, the difference is very small and we should still be able to extract the triples, despite the existence of noise peaks and missing peaks.

At the same time, adjacency information is extracted from the three spectra. The connection between spin systems from adjacent amino acid residues will be set up by connectivity information from CB and CA peaks. CB_i, CA_i, CB_{i-1}, and CA_{i-1} are identified from spectra HNCACB and CBCAcoNH. CB_i,

CA_i, CB_{i-1}, and CA_{i-1} are recognized from each plane of the spectrum HN-CACB, each of which has the same NH chemical shift value. Two planes from the spectrum CBCAcoNH containing CB_i, CA_i and CB_{i-1}, CA_{i-1}, respectively, are used to verify the four peaks, and the adjacent pairs of spin systems are generated by this process. Including the signature information obtained from BMRB database, the information needed for the assignment process is ready.

Further, a new computational model is proposed from our group (Dr. Lin's group at the University of Alberta) to deal with more complicated cases of peak grouping. In this model, peak grouping is formulated as a weighted bipartite matching problem, where two sides of vertices represent the peaks from two distinct spectra respectively, and the weight of edge represents the probability of mapping them to a common amino acid residue, taken as the square root of the product of the differences of HN chemical shifts and NH chemical shifts. The solution for this model is to find a minimum weight matching which indicates how two spectra can be merged into one. The resulting "super" spectrum from this repeating process represents the list of spin systems. Our approach is a globally optimal peak grouping method, without considering the match tolerance and the resolution of spectral data.

2.5.2 Assignment Starting with Spin Systems

This type of assignment method assigns the spin system right after peak grouping, without any adjacency information. Instead of determining the adjacency information during grouping, adjacency information determination is combined into spin systems assignment. A few recent works utilize this scheme: PACES [15], AutoAssign [53], and CISA [43].

PACES represents the adjacency relationships among spin systems as a directed network, and enumerates all the possible paths in it. Each such path represents a possible chain of the spin systems which might be mapped to a segment of the polypeptide chain of the target protein. PACES validates each path by mapping it to the most likely segment associated with the spin system signature information.

AutoAssign combines adjacency determination and the assignment processes, and uses them to validate each other. This combination dramatically reduces the total number of possible paths comparing to PACES. AutoAssign keeps a list of spin systems that nuclei in this amino acid residue may generate. For each pair of spin systems, AutoAssign checks if they could be mapped in the two lists of spin systems from two adjacent residues, respectively. If they can, then the pair is considered as a valid adjacent pair. However, the extension of the spin system path and the ambiguities of the adjacencies among the spin systems increase the complexity of this approach. In practice, AutoAssign requires extra information to reduce the complexity.

CISA [43] proposes another way to combine adjacency determination and peak assignment. The algorithm employs a *best-first* search incorporated with many other heuristics. A string of connected spin systems typically has a much better score at the correct mapping position in the target protein sequence than almost all the other (incorrect) mapping positions, especially when the string are long. Therefore, a string of spin systems having a higher mapping score are more likely to be correct. In other words, adjacency and assignment support each other. CISA starts with an *Open List* of strings and seeks to expand the string with the best mapping score. The succeedingly descendant strings are appended to Open List only if their normalized mapping scores are better than their ancestor's. Another list, *Complete List*, saves strings which are not further expandable. Once Open List becomes empty, high confident strings with their mapping positions are filtered out from the Complete List with the conflicts resolved in a greedy fashion. The preliminary simulation results in [43] show that CISA outperforms PACES, AutoAssign, and Random Graph Approach significantly, and many instances could not be solved by them can be solved by CISA.

2.5.3 Assignment Starting with Spin Systems and Adjacency Constraint

In this type of method, peak grouping and adjacency information determination are achieved at the same time. In other words, spin systems are connected

into small spin system chains by adjacency information, which must be mapped to a polypeptide segment on the target protein sequence [52].

MAPPER is a semi-automatic NMR assignment program that also starts with spin systems and their adjacencies, but proceeds with the assignment in a different manner. In more detail, the input to the program consists of the target protein sequence, the spectroscopically assembled short fragments of sequential connected residues, and CA and CB chemical shifts or amino acid type information for each spin system. MAPPER performs first an individual mapping to enumerate all the possible mappings for each fragment, and then performs an exhaustive search for global mapping (i.e., self-consistent mappings of all fragments) to obtain an unambiguous assignment. The global mapping is performed by fragment-nested loops, and the forbidden branches of the search tree will be cut as early as possible during the search. The only permissible overlap in global mapping is the overlap between two fragments which share one common residue, since the corresponding chemical shift values for the endpoint atoms satisfy a user-defined tolerance.

The assignment algorithm developed by our group uses the *Constrained Bipartite Matching* (CBM) formulation. The CBM model is basically the same as the normal weighted bipartite graph matching. The differences are the group of amino acid residues are ordered as their linear order in the target protein sequence, and the group of spin systems are partitioned into subsets, spin system strings, which must be mapped to a segment in the target protein sequence [52]. In theory, CBM problem is NP-hard even when the bipartite graph is complete [52]. The algorithm can be described as a two-phase procedure: the first phase is a *greedy filtering* procedure in which a certain number of best possible mappings are selected for the identified strings; the second phase is a maximum weight bipartite matching procedure in which the mapping between the isolated spin systems and the rest of the residues for every combination of string mappings is accomplished. The algorithm reports the best assignment from all combinations in terms of the assignment confidence — the score. This algorithm automates the assignment process at a global view, which produce an assignment within seconds on a Pentium IV PC.

2.5.4 Assignment Starting with Peak Lists

Most automated peak assignment programs apply the same general strategies as described above, to perform peak grouping, adjacency determination and assignment. However, the assignment methods starting with peak lists achieve peak grouping, adjacency determination, and assignment at the same time. Therefore, ambiguities arising at each step could be generally resolved in this way. If such ambiguities still couldn't be resolved at that moment, then manual adjustments have to be done. The following is a brief description of the proposed system [8, 9].

(1) All peaks from input spectra are put together to form a super peak list, where suitable shuffling is required to ensure the spectra have the same reference point (NH and HN chemical shifts are employed); (2) A clustering algorithm is applied on the super peak list to generate peak clusters such that peaks within a cluster share close NH and HN chemical shifts, where the number of clusters is set to the estimated number of spin systems using the target protein sequence. (Note that some different amino acid residues might have close NH and HN chemical shifts and thus multiple spin systems might reside in a cluster); (3) Since we cannot distinguish inter-residue and intra-residue peaks, an *undirected* graph $G = (V, E)$ is defined where each vertex represents a cluster, and two vertices are adjacent if they contain close chemical shifts for some nuclei. (Excluding NH and HN, tolerance thresholds are set); (4) A *best-first* search algorithm is applied which takes in the score scheme determined in the above to find a path cover for graph G. At the same time, the direction of a path will be determined using the spectral nature, with the exception that when the direction cannot be determined, then two directed copies of it are generated. The output of the search algorithm is a (directed) path cover of G with their mapping positions to the target protein sequence.

We note that such a system has a strong capability in resolving ambiguities and in cross-validation. An existing assignment algorithm GARANT [8, 9] is the most similar to the proposed system. GARANT starts with peak lists

in two dimensional COSY and two-dimensional NOESY spectra, and uses the knowledge of magnetization transfer pathways as the input. It represents peak assignment as an optimal match between two graphs, of which one is built for expected peaks predicted by combining knowledge of the primary structure and the magnetization transfer paths, and the other is for the observed peaks. It employs a genetic algorithm combined with a local optimization routine to find the optimal homomorphism of the graphs of the expected and observed peaks; this is evaluated by a sophisticated statistical score scheme based on *mutual information*.

2.5.5 Scoring Scheme

Scoring schemes are used to measure the match between a group of spin systems and a peptide of amino acid residues in the target protein. By including various terms, the scoring scheme will be able to give the probability of the accuracy of each match between two groups. There are three representative scoring schemes which are histogram-based score learning [44], representative-based score learning, and multiclass SVM with error-correcting output codes.

Histogram-based score learning [44] includes terms representing amino acid type, secondary structure type, and chemical shift values. For example, the score of the match between the spin system (HN_i, NH_i, CA_i, CB_i) and a specific amino acid type aa with a specific secondary structure ss, is related to the probabilities of occurrences of chemical shift values NH_i, CA_i and CB_i with amino acid type aa and secondary structure ss [44]:

$$\begin{aligned} & \text{score}((\text{HN}_i, \text{NH}_i, \text{CA}_i, \text{CB}_i) \mid (aa, ss)) \\ = \; & 10 \times \log \Big(\text{Prob}(aa, ss, \text{NH}_i) \times \text{Prob}(aa, ss, \text{CA}_i) \times \text{Prob}(aa, ss, \text{CB}_i) \Big), \end{aligned}$$

where

$$\text{Prob}(aa, ss, \text{XX}_i) = \frac{N(aa, ss, \text{XX}_i)}{N(aa, ss)}.$$

$N(aa, ss, \text{XX}_i)$ and $N(aa, ss)$ are numbers of entries in BMRB database with (aa, ss, XX_i) and (aa, ss).

Representative-based score learning applies clustering methods from data mining. Clustering is an unsupervised learning, but with the number of classes

pre-specified. One of the most popular algorithms is *Ordering Points to Identify the Clustering Structure* (OPTICS) [6] which is a hierarchical density-based clustering method for computing an augmented clustering ordering of objects for automatic and interactive clustering analysis. In this scoring scheme, each pair (aa, ss) represents a class. The probability for each spin system v generated by nuclei with (aa, ss) is measured by the sum of the Euclidean distance between the spin system v and the j-th representative spin system v_j for combination (aa, ss) as shown in Equation (2.5):

$$score(v \mid (aa, ss)) = \sum_{j=1}^{d} ||v, v_j||_2, \tag{2.5}$$

where $v1$, v_2, ..., v_d is d representative spin systems from the training set. The score represents the Euclidean distance between two spin systems, so the less the score, the more likely the spin system is generated by amino acid residue with the combination.

The third score scheme learning employs the multiclass SVM with error-correcting output codes. The code matrix is 60 × 64 which is generated by using the *Randomized Hill Climbing* algorithm [19]. Each row represents a combination of amino acid and secondary structure, which is a 64-bit code. These 64 SVMs are trained and each of them produces entries in one column in the code matrix. Given a new spin system $v = (\mathrm{HN}_i, \mathrm{NH}_i, \mathrm{CA}_i, \mathrm{CB}_i)$, a 64-bit string is produced by running these 64 SVMs. The hamming distance between it and each of the 60 class strings is taken as the score of mapping v to the combination. Again, such a score measures the "distance" rather than likelihood and thus the lower the score, the more likely the spin system is generated by nuclei from the combination.

2.6 Structure Determination

This section discusses the procedures for determining protein secondary and three-dimensional structures by structural constraints. Section 2.6.1 introduces how to extract all needed structural information. Then, the following

two sections present the procedures of secondary and three-dimensional structure determination.

2.6.1 Structural Constraints Extraction

Three major types of structural constraints can be extracted from experimental data: distance constraints, torsion angle constraints, and orientational constraints. These constraints come mainly from scalar coupling, dipolar coupling, hydrogen bond, and torsion angle interactions which are particularly sensitive to 3D spatial molecular conformation. Chemical shifts derived from the peak assignment procedure are the primary information used for secondary structure determination.

NOE-derived Distance Constraints

Nuclear Overhauser Effect (NOE) is a common phenomenon for pairs of nuclei of any type, with spatial distance between them shorter than 5Å. NOE-derived distance constraints are the most important source of structural information in protein structure determination. In NOESY spectrum, NOE interactions between pairs of nuclei are shown by NOE peaks. Each dimension of the spectrum is represented by a chemical shift. For example, if there is a peak at the point (4.5ppm, 4.6ppm) of a 2D NOESY spectrum, an NOE interaction between an hydrogen atom with chemical shift 4.5ppm and an hydrogen atom with chemical shift 4.6ppm exists. The intensity of the NOE is proportional to the product of the inverse sixth power of the internuclear distance d_{ij} between nuclei i and j and a correlation function $f(\tau_c)$, as

$$NOE_{ij} \propto \frac{1}{(d_{ij})^6} f(\tau_c).$$

The structural information comes mainly from NOEs between two hydrogen atoms, especially from pairs of hydrogen atoms close in space but far away on the polypeptide chain. NOE peak intensities are commonly classified into *very weak*, *weak*, *medium*, *strong*, and *very strong* [40]. Each class is constrained by a

Table 2.1: Distance bounds for different NOE intensity classes

NOE intensity classes	distance [Å]	upper bound [Å]
very strong	2.3	2.5
strong	2.8	3.1
medium	3.1	3.4
weak	3.5	3.9
very weak	4.2	5.0

set of approximate distance bounds, as shown in Table 2.1. The lower bounds are often set to the sum of the van der Waals radii of the two protons. In order to solve the ambiguities arising from spin diffusion and spectral overlap, the concept of ambiguous distance constraint has been introduced, and is represented by

$$\bar{d}_{F_1,F_2} = \Big(\sum_{k=1}^{N(F_1,F_2)} d_k^{-6} \Big)^{-1/6},$$

in which k runs through all $N(F_1, F_2)$ contributions to a crosspeak at frequencies F_1 and F_2, and d_k is the internuclear distance between two protons with the contributions k determined by the coordinates of a model structure. All the distance constraints derived from NOEs can be incorporated into structure calculation models directly, such as distance geometry ([30], [11]) and torsion angle dynamics [27].

Secondary structure identification can be obtained by an empirical approach based on NOE information [51, 50]. An example is provided in Figure 2.3.

Hydrogen Bond Constraints

Hydrogen bond constraints are incorporated into the structure calculation as distance constraints. They are useful for preliminary protein structure calculation of larger proteins when NOE data is scarce. The hydrogen bond limits the acceptor(O)-donor(H) distance and the distance between acceptor(O) and the nitrogen atom(N), which is covalent to the donor(H) to the ranges [1.8Å, 2.0Å] and [2.7Å, 3.0Å], respectively, as shown in Figure 2.4(a).

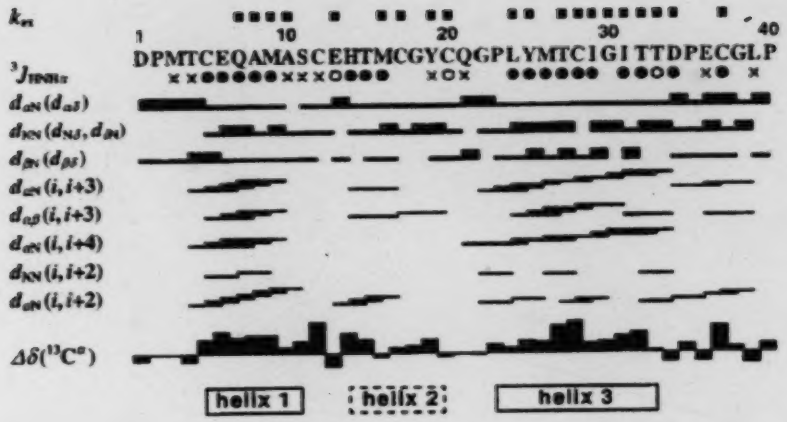

Figure 2.3: Example showing methods recommended for presenting NMR data supporting the secondary structure identification in proteins. The 40-residue protein, pheromone Er-2, is used as an illustration [41]. Above the amino acid sequence, black squares identify residues with observably slow hydrogen-exchange rates, k_{ex}, at the backbone amide (the conditions of the exchange experiment should be specified). Below the amino acid sequence, filled circles identify residues with ${}^3J_{\mathrm{HNH}\alpha} < 6.0$Hz, indicative of local α-type conformation; open circles correspond to ${}^3J_{\mathrm{HNH}\alpha} > 8.0$Hz, indicative of residues in extended chain conformation; crosses identify residues with ${}^3J_{\mathrm{HNH}\alpha}$ values 6.0 to 8.0 Hz. For the sequential proton-proton NOE connectivities, $d_{\alpha N}$, $d_{\delta N}$, $d_{\beta N}$ for Pro-Xxx dipeptides), thick and thin bars indicate strong and weak NOE intensities, respectively. The observed medium-range NOEs $d_{\alpha N}(i, i+3)$, $d_{\alpha\beta}(i, i+3)$, $d_{\alpha N}(i, i+4)$, $d_{NN}(i, i+2)$, and $d_{\alpha N}(i, i+2)$ are indicated by lines connecting the two residues that are related by the NOE. ${}^{13}C^{\alpha}$ chemical shifts relative to the random coil values, $\Delta\delta({}^{13}C^{\alpha})$, are plotted at the bottom of the figure, where positive values are shifts to lower field. The sequence locations of three helices are indicated at the bottom; broken lines are used to indicate that the identification of helix 2 from these data is uncertain.

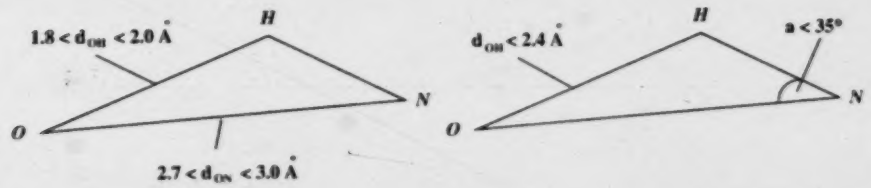

Figure 2.4: (a) Hydrogen bond restraints used during a structure calculation [47]. (b) Criterion used to detect hydrogen bonds when analyzing a structure [10, 36].

Hydrogen bond constraints have also been found useful in determining secondary structure elements such as α-helices and β-sheets. Hydrogen bonds can be detected through ^{3h}J (three covalent bonds) and ^{2h}J (two covalent bonds) scalar couplings [16, 20]. Furthermore, hydrogen bonds are inferable in α-helices, but not in β-sheets. For small proteins, heteronuclear $^{3h}J_{NC'}$ [16] and $^{2h}J_{HC'}$ [17] couplings indicate the presence of a hydrogen bond.

Torsion Angle Constraints

Torsion angle constraints are obtained from the three-bond vicinal coupling constant ^3J by the Karplus equation:

$$^3J(\theta) = A\cos^2\theta + B\cos\theta + C, \tag{2.6}$$

where A, B and C are empirical parameters determined by a best fit of the measured J values to the corresponding values calculated with Equation (2.6). The most commonly used Karplus parameters are listed in Table 2.2. θ is the torsion angle between the four atoms connected by three vicinal bonds. Therefore, given ^3J-coupling constants, allowed ranges of the relevant torsion angles can be derived. The main contributions of torsion angle constraints are to provide the local conformation. Although they do confine backbone torsion angles ϕ and ψ, and the side chain χ torsion angles, their contributions to the global fold are limited.

RDCs Orientational Constraints

Residual dipolar couplings (RDCs) constraints are introduced into structure calculation as orientational constraints. Structural information is obtained from RDCs by observing internuclear dipolar interactions. In solution, molecules are isotropically oriented, so the internuclear dipolar interactions average to zero and cannot be observed. If proteins are immersed into an anisotropic environment which has a different property according to different direction, such as solutions containing phages or bicelles, dipolar couplings no longer average to zero and produce an observable residual dipolar coupling. The residual

Table 2.2: Karplus relations, ${}^3J(\theta) = A\cos^2\theta + B\cos\theta + C$, for proteins between a vicinal scalar coupling constant ${}^3J$ and the corresponding torsion angle θ, defined by the three covalent bonds between the two scalar coupled atoms. "Offset" in the table represents the difference between θ and the standard torsion angle ϕ, ψ or χ^1. In the case of β-methylene protons, the first number is for $H^{\beta 2}$, the second for $H^{\beta 3}$.

Angle	Coupling	A (Hz)	B (Hz)	C (Hz)	Offset (degrees)	Reference
ϕ	$H^N - H^\alpha$	6.98	-1.38	1.72	-60	[46]
	$H^N - C'$	4.32	0.84	0.00	180	[46]
	$H^N - C^\beta$	3.39	-0.94	0.07	60	[46]
	$C'_{i-1} - H^\alpha$	3.75	2.19	1.28	120	[46]
	$C'_{i-1} - C^\beta$	1.59	-0.67	0.27	-120	[33]
ψ	$H^\alpha - N_{i+1}$	-0.88	-0.61	-0.27	-120	[45]
χ^1	$H^\alpha - H^\beta$	9.50	-1.60	1.80	-120/0	[39]
	$N - H^\beta$	-4.40	1.20	0.10	120/-120	[38]
	$C' - H^\beta$	7.20	-2.04	0.60	0/120	[21]

dipolar coupling (D) between two nuclei i and j is calculated by

$$D_{ij}(\theta,\phi) = D_a\left[(3\cos^2\theta - 1) + \frac{3}{2}D_r\sin^2\theta\cos 2\phi\right],$$

where D_a is the dipolar coupling tensor and D_r is the rhombicity. θ and ϕ are cylindrical coordinates describing the orientation of the internuclear vector $\vec{ij}$ in the principal axis system of the molecular alignment tensor. So, given the molecular alignment tensor, RDCs provide the orientation of internuclear vectors relative to an external reference frame, which is defined in the structure calculation process as an orthogonal axis system [42]. Orientational constraints derived from RDCs can be included in the structure calculation process as a pseudo-potential energy term, similar to other constraints.

2.6.2 Secondary Structure Determination

The empirical correlation between protein secondary structure and chemical shifts of C^α, C^β, C', H^α and N has been found, and a number of methods make use of this correlation to predict a reliable protein secondary structure. CSI

[48, 49] and TALOS [18] are two popular ones. In the rest of this subsection, there is an introduction of those two methods.

CSI

The library referred to as Chemical Shift Index [48] includes chemical shifts of atoms C^{α}, C^{β}, C^{ι}, H^{α} and N. The main principle of CSI is the comparison of the chemical shifts obtained from experiments with the Chemical Shift Index. CSI is able to predict the protein secondary structure type by comparing the chemical shifts from the protein sequence and those from the Chemical Shift Index.

As shown in Table 2.3, all the CSI entries for the chemical shifts in the CSI library are in the form of ranges. If the chemical shift of amino acid residue on the sequence is greater than the range shown in the table, this residue is marked +1. If the chemical shift of amino acid residue on the sequence is smaller than the range, the residue is marked -1. If the chemical shift of amino acid residue on the sequence is within this range, the residue is marked 0. The secondary structure element can be roughly recognized from the marks of each amino acid residue. The rules of that are as follows: (1) any segment with 4 "-1's" without interrupting by a "+1" is a helix; (2) any segment with 3 "+1's" without interrupting by a "-1" is a strand; (3) any other combination is a coil; (4) end points of helices and strands can be recognized by the first occurrence of an opposite mark or two consecutive 0's in the CSI; (5) the local density of "-1's" and "+1's" measured for a window of 4 to 5 residues has to exceed 70% to define a structured element. After a specific secondary structure element is recognized, the Chemical Shift Index will give out the angle restraints of torsion angles involved in the protein local conformation. The resulting information can be used as structural constraints for structure calculation.

TALOS

TALOS [18] not only considers the information from chemical shift, but also takes sequence similarity into account. This method will give the user the most

Table 2.3: CSI entries for the H^{α} chemical shifts [48]

residue	$\alpha - {}^1H$ range (ppm)	residue	$\alpha - {}^1H$ range (ppm)
Ala	4.35±0.10	Met	4.52±0.10
Cys	4.65±0.10	Asn	4.75±0.10
Asp	4.76±0.10	Pro	4.44±0.10
Glu	4.29±0.10	Gln	4.37±0.10
Phe	4.66±0.10	Arg	4.38±0.10
Gly	3.97±0.10	Ser	4.50±0.10
His	4.63±0.10	Thr	4.35±0.10
Ile	3.95±0.10	Val	3.95±0.10
Lys	4.36±0.10	Trp	4.70±0.10
Leu	4.17±0.10	Tyr	4.60±0.10

similar 10 triplets in terms of the similarity between sequences in the database and the query sequence. If the central residues in the given 10 triplets show similar backbone angles, the average of them will be considered as the angular restaints for the query sequence. The diagram of the whole procedure of TALOS is shown in Figure 2.5.

The database used in TALOS includes C^{α}, C^{β}, C', H^{α} and N chemical shifts for 20 proteins with a high resolution X-ray structure. TALOS utilizes both similarity of sequence and chemical shift to predict the restraints for backbone angles. The idea is based on the assumption that if the adjacent amino acid residues have similar chemical shifts to a string of amino acid residues in the database, the central amino acid residues in two strings should have similar backbone torsion angles. The similarity in the amino acid residue types between two strings can then be used as a complement for the criterion.

TALOS uses the NMRWish(a companion package to the NMR data processing and analysis tool–NMRPipe), which is written in the language Tcl/Tk. It provides a formula to measure the similarity between the central amino acid residue of two strings, known as similarity factor $S(i,j)$ defined by:

$$\begin{aligned} S(i,j) &= \sum_{n=-1}^{+1} [k_n^0 \Delta^2_{ResType} + k_n^1(\Delta\delta C^{\alpha}_{i+n} - \Delta\delta C^{\alpha}_{j+n})^2 \\ &\quad + k_n^2(\Delta\delta N_{i+n} - \Delta\delta N_{j+n})^2 + k_n^3(\Delta\delta C^{\beta}_{i+n} - \Delta\delta C^{\beta}_{j+n})^2 \end{aligned}$$

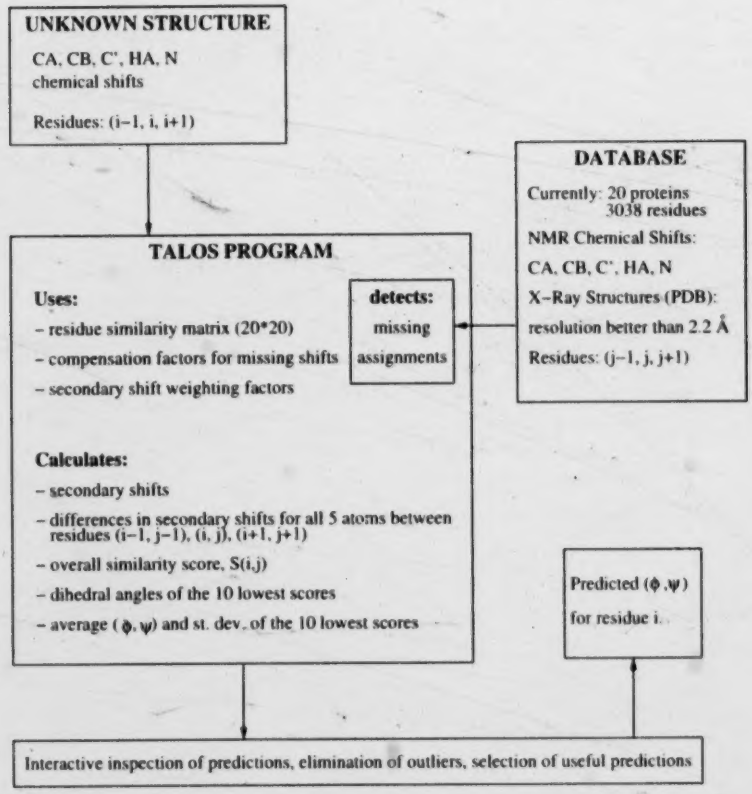
UNKNOWN STRUCTURE
CA, CB, C', HA, N
chemical shifts
Residues: (i−1, i, i+1)
DATABASE
Currently: 20 proteins
3038 residues
NMR Chemical Shifts:
CA, CB, C', HA, N
X−Ray Structures (PDB):
resolution better than 2.2 Å
Residues: (j−1, j, j+1)
TALOS PROGRAM
Uses:
– residue similarity matrix (20*20)
– compensation factors for missing shifts
– secondary shift weighting factors
detects:
missing
assignments
Calculates:
– secondary shifts
– differences in secondary shifts for all 5 atoms between residues (i−1, j−1), (i, j), (i+1, j+1)
– overall similarity score, S(i,j)
– dihedral angles of the 10 lowest scores
– average (ϕ, ψ) and st. dev. of the 10 lowest scores
Predicted (ϕ, ψ)
for residue i.
Interactive inspection of predictions, elimination of outliers, selection of useful predictions

Figure 2.5: TALOS flowchart [18]

Table 2.4: Residue similarity factors, $\Delta_{ResType}$ [18]

Res.	A	R	N D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W Y	V
A	0	1	1	1	1	1	2	1	2	1	1	1	2	3	1	2	2	2
R	1	0	1	1	1	1	2	1	2	1	0	1	1	3	1	2	1	2
N/D	1	1	0	1	1	1	2	1	2	1	1	1	1	3	1	2	1	2
C	1	1	1	0	1	1	2	1	2	1	1	1	1	3	1	2	1	2
Q	1	1	1	1	0	1	2	1	2	1	1	1	1	3	1	2	1	2
E	1	1	1	1	1	0	2	1	2	2	2	1	1	3	1	2	1	2
G	2	2	2	2	2	2	0	3	3	3	3	3	3	3	3	3	3	3
H	1	1	1	1	1	1	3	0	2	1	2	2	1	3	2	2	1	2
I	2	2	2	2	2	2	3	2	0	1	2	2	2	3	2	1	2	0
L	1	1	1	1	1	2	3	1	1	0	1	1	1	3	2	2	1	2
K	1	0	1	1	1	2	3	2	2	1	0	1	2	3	1	2	2	2
M	1	1	1	1	1	1	3	2	2	1	1	0	2	3	1	2	2	2
F	2	1	1	1	1	1	3	1	2	1	2	2	0	3	2	2	0	1
P	3	3	3	3	3	3	3	3	3	3	3	3	3	0	3	3	3	3
S	1	1	1	1	1	1	3	2	2	2	1	1	2	3	0	1	2	2
T	2	2	2	2	2	2	3	2	1	2	2	2	2	3	1	0	1	1
W/Y	2	1	1	1	1	1	3	1	2	1	2	2	0	3	2	1	0	1
V	2	2	2	2	2	2	3	2	0	2	2	2	1	3	2	1	1	0

$$+k_n^4(\Delta\delta C'_{i+n} - \Delta\delta C'_{j+n})^2 + k_n^5(\Delta\delta H^{\alpha}_{i+n} - \Delta\delta H^{\alpha}_{j+n})^2],$$

where k_n are empirically optimized factors, and $\Delta\delta N_{i+n}$ represents the chemical shift of nitrogen atom on $(i+n)$th residue. Two amino acid residues are more similar if the similarity factor $S(i,j)$ is smaller. The TALOS program gives out 10 strings with the smallest similarity factors. The similarity of amino acid residue type is evaluated by a 20×20 matrix shown in Table 2.4.

2.6.3 3D Structure Calculation

As described in the previous section, we can obtain different types of constraints, such as interproton distance and dihedral angle constraints. The problem discussed in this section will concern the structure calculation algorithms based on these experimentally derived constraints. In this section, four

kinds of algorithms which are the most widely used in structure calculation domain are introduced.

Metric Matrix Distance Geometry

Distance geometry is the earliest algorithm used in structure calculation. The inherent principle of distance geometry is that it is possible to calculate Cartesian coordinates of a set of points if all the distances among those points are known.

Metric matrix distance geometry [25] uses an $N \times N$ matrix G to solve the problem, in which N is the number of atoms on the protein sequence. Metric matrix G has the property by definition that it has only three positive eigenvalues. Other $N-3$ eigenvalues are all equal to zero. The matrix G_{ij} is constructed as follows:

$$G_{ij} = r_i \cdot r_j = \left\{ \begin{array}{ll} \frac{1}{N}\sum_{k=1}^{N} D_{ik}^2 - \frac{1}{2N^2}\sum_{k,l=1}^{N} D_{kl}^2 & \text{if } i = j \\ \frac{D_{ij}^2 - G_{ii} - G_{jj}}{2}. & \text{if } i \neq j \end{array} \right\},$$

where $N \times N$ distance matrix D_{ij} is calculated by

$$D_{ij} = |r_i - r_j|,$$

where r_i $(i = 1, 2, ..., N)$ denotes the coordinate of atom i in Cartesian three-dimensional space. Nevertheless, in practice, the metric matrix generated from experimentally derived constraints does not possess three positive eigenvalues under some circumstances. In order to get the metric matrix we need, a series of triangle inequalities check needs to be accomplished [11].

The coordinates of all the atoms in Cartesian three-dimensional space can be calculated by

$$r_i^\alpha = \sqrt{\lambda^\alpha} e_i^\alpha \qquad (i = 1, 2, ..., N; \alpha = 1, 2, 3),$$

where λ^α and e^α denote eigenvalues and eigenvectors of the matrix G. r_i^1, r_i^2 and r_i^3 represent the coordinate of atom i in x, y and z axis, respectively.

Based on the introduction above, we can calculate the coordinates of all the atoms if we know the exact distances. However, not all the distance constraints are available; some of these constraints are in the form of a range. At first,

those exact values of the distances derived from covalent structure can be input into the matrix directly. Then, for those distances with no information from experiments, a large value will be set to be the upper bound, e.g., 40Å, and the lower bound will be zero. Finally, a random value will be picked between the lower bound and the upper bound of the range to feed into the matrix. Repeating this procedure, we can obtain a group of conformers which build the conformational space. The conformers obtained from this algorithm may be seriously distorted. Hence, they needs to be refined by other methods, such as conjugate gradient minimization [30].

This algorithm is no longer used in the protein structure calculation, and better algorithms will be introduced in the following sections. However, the conformers generated by this algorithm could be used as a initial structure for other algorithms.

Variable Target Function Method

The structure calculation problem in the variable target function method is formulated as a target function minimization problem. The target function represents the constraint violations, and has the following properties: (1) target function $T(\phi_1, ..., \phi_n)$ is equal to zero if all the experimentally derived constraints are satisfied; (2) $T(\phi_1, ..., \phi_n) \leqslant T(\theta_1, ..., \theta_n)$ if $(\phi_1, ..., \phi_n)$ satisfies the constraints better than $(\theta_1, ..., \theta_n)$. In addition, the variable target function method is an algorithm in torsion angle space. The degrees of freedom are n torsion angles (ϕ_1, ϕ_2, ..., ϕ_n).

In order to reduce the danger of being trapped in a local minimum, variable target function increases "target size" in a step-wise fashion. The target size progresses from the intra-residue level up to the whole polypeptide chain (Figure 2.6). Therefore, the local conformation of the protein sequence will be obtained first, and the global fold of the protein sequence can only be established near the end of the calculation.

The target functions implemented in variable programs are somewhat diverse. The one implemented in the program DIANA [26], which is discussed in detail here, will be compared with the target function implemented in the

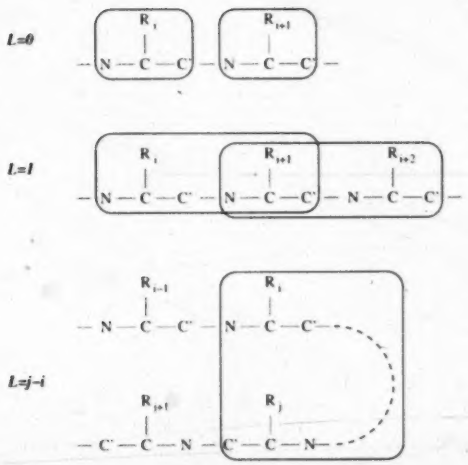

Figure 2.6: Active restraints at various minimization levels L of the variable target function algorithm. At a given minimization level L, all distance restraints between residues i and j with $|j-i| \leqslant L$ are considered [25].

program DISMAN [12].

In the program DIANA, it is assumed that there exist n_u distance upper bounds, n_l distance lower bounds, n_v van der Waals' repulsion lower bounds, and n_a dihedral angle constraints. Thus, the upper or lower bound b of the distance between atom α and atom β is represented by:

$$\begin{array}{ll} (\alpha_i^u, \beta_i^u, b_i^u), & i = 1, ..., n_u; \\ (\alpha_i^l, \beta_i^l, b_i^l), & i = 1, ..., n_l; \\ (\alpha_i^v, \beta_i^v, b_i^v), & i = 1, ..., n_v; \\ (a_i, \phi_i^{\min}, \phi_i^{\max}), & i = 1, ..., n_a. \end{array}$$

The target function is constructed as

$$T = \sum_{c=u,l,v} w_c \sum_{i \in I_c} \Big(\Theta_c \Big(\frac{d_i^{c2} - b_i^{c2}}{2 b_i^c} \Big) \Big)^2 + w_a \sum_{i=1}^{n_a} \Big(1 - \frac{1}{2} \Big(\frac{\Delta_i}{\Gamma_i} \Big)^2 \Big) \Delta_i^2,$$

with

$$\Theta_c(t) = \left\{ \begin{array}{ll} \max(0,t), & \text{if } c = u; \\ \min(0,t), & \text{if } c = l, v. \end{array} \right\},$$

where, w_c and w_a are weight factors for the corresponding items; $I_c \in \{1, ..., n_c\}$ with $c = u, l, v$, d_i^c represents the distance between atom α_i^c and atom β_i^c. Δ, denoting the signed dihedral angle constraint violation, is calculated as

$$\Delta = \left\{ \begin{array}{ll} 0, & \text{if } \phi_a \in [\phi^{\min}, \phi^{\max}]; \\ -\Delta^{\min}, & \text{if } \phi_a \notin [\phi^{\min}, \phi^{\max}] \text{ and } \Delta^{\min} \leqslant \Delta^{\max}; \\ \Delta^{\max}, & \text{if } \phi_a \notin [\phi^{\min}, \phi^{\max}] \text{ and } \Delta^{\min} > \Delta^{\max}; \end{array} \right. \qquad (2.7)$$

with

$$\Delta^{\min} = \min\{|\hat{\phi}^{\min} - \hat{\phi}_a|, 2\pi - |\hat{\phi}^{\min} - \hat{\phi}_a|\}, \tag{2.8}$$

and

$$\Delta^{\max} = \min\{|\hat{\phi}^{\max} - \hat{\phi}_a|, 2\pi - |\hat{\phi}^{\max} - \hat{\phi}_a|\}. \tag{2.9}$$

Here, $\hat{\phi}$ denotes the equivalent value of ϕ in the interval $[0, 2\pi]$, i.e.,

$$\hat{\phi} = \phi + n \cdot 2\pi, \qquad n \in \mathbf{Z}.$$

Γ denotes the half-width of the forbidden interval of dihedral angle values:

$$\Gamma = \pi - \frac{\phi^{\max} - \phi^{\min}}{2}.$$

The target function T above is the one implemented in the program DIANA. However, the target function T' implemented in the program DISMAN is somewhat different in the treatments of steric constraints and dihedral angle constraints than function T, and is calculated as

$$T' = \sum_{c=u,l} w_c \sum_{i \in I_c} \left(\Theta_c\left(\frac{d_i^{c2} - b_i^{c2}}{2b_i^c}\right)\right)^2 + \frac{w_v}{4} \sum_{i \in I_v} (\Theta_v(d_i^{v2} - b_i^{v2}))^2 + 4w_a \sum_{i=1}^{n_a} \left(1 - \frac{1}{2}\frac{|\Delta_i|}{\Gamma_i}\right)^2 \left(\frac{\Delta_i}{\Gamma_i}\right)^2.$$

The drawback of the implementation of variable target function is that it is still possible to be trapped in local minima [12]. Also, because of the low success rate, a large number of randomized start structures need to be used in the calculations in order to generate a group of good conformers. On the other hand, compromise may be made between small constraints violation and computational complexity. The variable target function method has been proved to achieve better results in determining α-helical proteins than β-proteins [28]. That may be attributed to the factor that β sheet has more complex topology which involves long-range distance constraints.

Molecular Dynamics in Cartesian Space

Molecular dynamics is a method for simulating the movement of a molecular system. Simulated annealing method simulates a slowly cooling process of a molecular system from an extremely high temperature. Therefore, the method combining molecular dynamics (MD) and simulated annealing (SA) is

called "molecular dynamics simulation". The distinctive feature of molecular dynamics simulation, compared to other target function minimization methods, is the presence of kinetic energy. The presence of kinetic energy greatly reduces the probability of being trapped in a local minimum. Different from the standard MD [23], molecular dynamics simulation uses a pseudo-potential energy function as the target function.

The remainder part of this section will introduce molecular dynamics simulation in Cartesian space. The degrees of freedom of this method are the Cartesian coordinates of the atoms.

The governing equation of the molecular system of n atoms for molecular dynamics simulation in Cartesian space is Newton's equation of motion:

$$m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{F}_i \qquad (i = 1, 2, ..., n),$$

where, the forces $\vec{F}_i$ are given by the negative gradient of the potential energy function E_{pot} in Cartesian space:

$$\vec{F}_i = -\nabla_i E_{pot}. \tag{2.10}$$

The potential function consists of the terms representing bond length and bond angle potentials from covalent structure, torsion angle potentials, distance constraints from non-bonded interaction, and other distance and angle constraints in the form of ranges. For example, the potential function in the program XPLOR [13] is

$$\begin{aligned} E_{pot} \;=\; & \sum_{\text{bonds}} k_b (r - r_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}} k_\phi (1 + \cos(n\phi + \delta)) \\ & + \sum_{\text{impropers}} k_\phi (\phi - \delta)^2 + \sum_{\text{nonbonded pairs}} k_{\text{repel}} (\max(0, (sR_{\min})^2 - R^2))^2 \\ & + \sum_{\text{distance restraints}} k_d \Delta_d^2 + \sum_{\text{angle restraints}} k_a \Delta_a^2, \end{aligned} \tag{2.11}$$

where k_b, k_θ, k_ϕ, k_{repel}, k_d and k_a denote a variety of force constants; r and θ are actual bond length and bond angle with r_0 and θ_0 corresponding to fixed bond length and bond angle from covalent structure; ϕ denotes the actual torsion angle or improper angle value; δ is the offset of the torsion angle or improper

potentials; R represents the actual distance between a non-bonded interaction atom pair; $R_{\min}$ is the distance when van der Waals potential between two atoms reaches its minimum; s is a scaling factor; and Δ_d and Δ_a are the distance and angle constraint violation, respectively.

In order to calculate the trajectory, Newton's equation of motion is numerically integrated by different integration schemes, such as 'leap-frog' integration [13]:

$$\vec{v}_i(t + \Delta t/2) = \vec{v}_i(t - \Delta t/2) + \Delta t \vec{F}_i(t)/m_i, \tag{2.12}$$

$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \Delta t \vec{v}_i(t + \Delta t/2). \tag{2.13}$$

The starting structure is provided by metric matrix distance geometry, so the initial coordinates r_i^0 are known, and the initial velocities are randomly assigned according to a Maxwell-Boltzmann distribution. The probability that atom i has the velocity v_i at temperature T is given by:

$$P(v_i) = \Big(\frac{m_i}{2\pi k_B T}\Big)^{1/2} \exp\Big[-\frac{1}{2}\frac{m_i v_i^2}{k_B T}\Big], \tag{2.14}$$

where $k_B = 1.38 \times 10^{-23}\mathrm{JK}^{-1}$ is the Boltzmann constant, and m_i is the mass of atom i. Given initial coordinates and velocities, it is possible to obtain the velocities and positions of atoms after time step Δt by Equation (2.12) and Equation (2.13). The magnitude of the time step Δt must be small enough to sample the fastest motions (of the order of $10^{-15}s$). During the procedure of molecular dynamics simulation, temperature control is one of the essential steps. Velocity Rescaling is commonly used to maintain the temperature. For each step of the slowly heating or cooling procedure, velocity needs to be adjusted by a scale constant to control the temperature. The simulated annealing protocol in the program X-PLOR includes adjustment of the steric repulsion.

The complete procedure can be described as follows: (1) obtain initial velocities from Equation (2.14) and initial structure from distance geometry; (2) calculate potential energy from Equation (2.11), then obtain forces from Equation (2.10); (3) determine new velocities after time step Δt from Equation (2.12) with known initial velocities and accelerations; (4) given initial positions

Table 2.5: Comparison of molecular dynamics simulations in Cartesian and torsion angle spaces

Quantity	Cartesian space	Torsion angle space
Degrees of freedom	$3N$ coordinates: $x_1, ..., x_N$	n torsion angles: $\theta_1, ..., \theta_n$
Equations of motion	Newton's equations: $m_i\ddot{x}_i = -\frac{\partial E_{pot}}{\partial x_i}$	Lagrange equations: $\frac{d}{dt}(\frac{\partial L}{\partial \dot{\theta}_k}) - \frac{\partial L}{\partial \theta_k} = 0$ $(L = E_{kin} - E_{pot})$
Kinetic energy	$E_{kin} = \frac{1}{2}\sum_{i=1}^{N} m_i \dot{x}_i^2$	$E_{kin} = \frac{1}{2}\sum_{k,l=1}^{n} M(\theta)_{kl}\dot{\theta}_k\dot{\theta}_l$
Mass matrix M	Diagonal, elements m_i	$n \times n$, non-diagonal, non-constant
Accelerations	$\ddot{x}_i = -\frac{1}{m_i}\frac{\partial E_{pot}}{\partial x_i}$	$\ddot{\theta} = M(\theta)^{-1}C(\theta, \dot{\theta})$ (n linear equations)
Computational complexity of acceleration calculation	Proportional to N	If solving system of linear equations: proportional to n^3 If exploiting tree structure of molecule: proportional to n

of atoms and current velocities, calculate new positions by 2.13; (5) repeat the above 4 steps, until the system is at equilibrium or the target temperature is reached. All 5 steps are repeated on every start structure to obtain a group of good conformers.

Torsion Angle Dynamics

The fundamental difference between torsion angle dynamics and molecular dynamics in Cartesian space is the degrees of freedom. Torsion angle dynamics use torsion angles as the degrees of freedom instead of Cartesian coordinates. As a representation, the program DYANA [27] is discussed in this section. A comparison between torsion angle dynamics and molecular dynamics in Cartesian space is made in Table 2.5.

Since the degrees of freedom are torsion angles, in the program DYANA, the molecular system of the protein has been represented as a tree structure with a fixed base rigid body and other n rigid bodies connected by n rotatable bonds, as illustrated in Figure 2.7. Each rigid body contains one or several atoms with unchangeable relative positions. The tree structure starts from the base rigid body, N-terminus of the polypeptide, and ends at the C-terminus

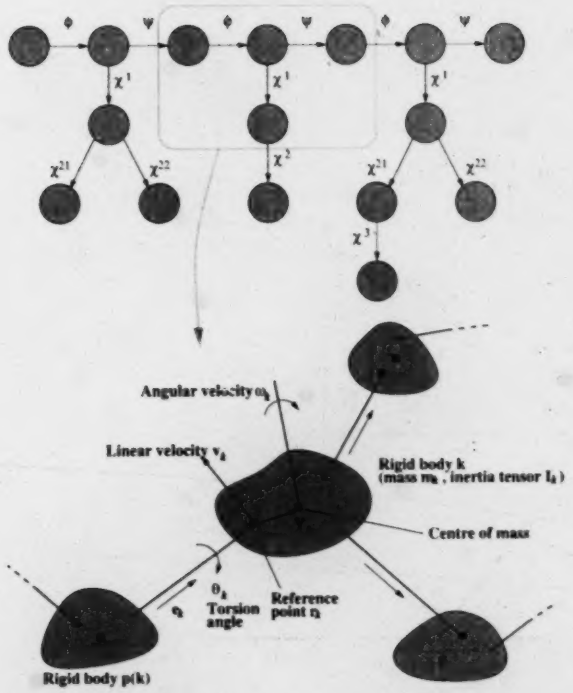

Figure 2.7: (a) Tree structure of torsion angles for the tripeptide Val-Ser-Ile. Circles represent rigid units. Rotatable bonds are indicated by arrows that point towards the part of the structure that is rotated if the corresponding dihedral angle is changed. (b) Excerpt from the tree structure formed by the torsion angles of a molecule, and various quantities required by the torsion angle dynamics algorithm of [34]

and the side-chains. Therefore, n torsion angles among $0, 1, ..., n$ rigid bodies are denoted by $\theta_1, \theta_2, ..., \theta_n$. In Figure 2.7, $p(k)$ denotes the preceding rigid body of rigid body k in the polypeptide chain. The torsion angle between two rigid bodies $p(k)$ and k is represented by θ_k. $\vec{e}_k$ denotes a unit vector illustrating the direction of the bond connecting rigid bodies $p(k)$ and k. $\vec{r}_k$ is a position vector of the 'reference point', i.e., the end point of the bond between $p(k)$ and k. The only movement allowed in this tree structure is the rotation of the bonds. The incompatible covalent structure of the tree structure, such as closed flexible rings, will be solved by the participation of other methods, such as molecular dynamics in Cartesian space.

In the program DYANA, the target function V represents potential energy E_{pot} by

$$E_{pot} = w_0 V,$$

in which, $w_0 = 10\text{kJmol}^{-1}\text{Å}^{-2}$ is an overall weight factor, and the target function itself is

$$V = \sum_{c=u,l,v} w_c \sum_{(\alpha,\beta)\in I_c} f_c(d_{\alpha\beta}, b_{\alpha\beta}) + w_d \sum_{k\in I_d} \Big(1 - \frac{1}{2}\Big(\frac{\Delta_k}{\Gamma_k}\Big)^2\Big)\Delta_k^2, \qquad (2.15)$$

where the function $f_c(d, b)$ may be in the form of any of the three equations below:

$$\begin{aligned} f_c(d,b) &= \Big(\frac{d^2 - b^2}{2b}\Big)^2, \\ f_c(d,b) &= (d-b)^2, \\ f_c(d,b) &= 2\beta^2 b^2\Big[\sqrt{1 + \Big(\frac{d-b}{\beta b}\Big)^2} - 1\Big]. \end{aligned}$$

This illustrates the effect of a violated distance constraint on the target function. In Equation (2.15), $b_{\alpha\beta}$ represents lower and upper bounds, while $d_{\alpha\beta}$ denotes the actual distance between atoms α and β. I_c $(c = u, l, v)$ are atom pairs (α, β) with upper and lower bounds, and van der Waals's repulsion distance bounds, respectively; I_d is the set of restrained torsion angles; w_c and w_d are all weighting factors in various constraints; and Δ_k is the size of the torsion angle constraints violation (refer to Equation (2.7–2.9)). Γ_k, half-width of the

forbidden range of torsion angle k, is

$$\Gamma_k = \pi - \frac{\theta_k^{max} - \theta_k^{min}}{2},$$

where θ_k^{min} and θ_k^{max} are lower and upper bound of torsion angle k, respectively.

As described in Figure 2.7, the angular and the linear velocity of rigid body k are relevant to the angular and the linear velocity of rigid body $p(k)$ and the effect of rotation of the bond between them. It is therefore easy to obtain the equations of angular velocity $\vec{w}_k$ and linear velocity $\vec{v}_k = \dot{\vec{r}}_k$ of the reference point for rigid body k with $k = 1, 2, ..., n$:

$$\vec{w}_k = \vec{w}_{p(k)} + \vec{e}_k \dot{\theta}_k, \tag{2.16}$$

$$\vec{v}_k = \vec{v}_{p(k)} - (\vec{r}_k - \vec{r}_{p(k)}) \wedge \vec{w}_{p(k)}. \tag{2.17}$$

In Equation (2.16), angular velocity w_k of the rigid body k is the sum of the angular velocities of the preceding rigid body $p(k)$ and the bond connecting them. Therefore, kinetic energy is given by

$$E_{\text{kin}} = \frac{1}{2} \sum_{k=1}^{n} [m_k \vec{v}_k^2 + \vec{w}_k \cdot \mathbf{I}_k \vec{w}_k + 2\vec{v}_k \cdot (\vec{w}_k \wedge m_k \vec{Y}_k)],$$

where $\vec{Y}_k$ denotes the vector from the reference point $\vec{r}_k$ to the center of mass. The mass is represented by m_k, and the inertia tensor by $\mathbf{I}_k$.

The inertia tensor of the rigid body k is a symmetric 3×3 matrix, given by

$$(\mathbf{I}_k)_{ij} = \sum_a m_a (\vec{y}_a^2 \delta_{ij} - y_{ai} y_{aj}),$$

where a is a variable going through all the atoms contained in the rigid body k, $\vec{y}_a$ is the vector from reference point to the atom a, and δ_{ij} is the Kronecker symbol. In practice, all the rigid bodies have been treated as solid spheres with mass m_k to improve the efficiency of the algorithm,

$$\begin{aligned} \vec{Y}_k &= \vec{0}, \\ \mathbf{I}_k &= \frac{2}{5} m_k \rho^2 \mathbf{1}_3, \end{aligned}$$

in which, ρ is the radius with center at reference point $\vec{r}_k$ and $\mathbf{1}_3$ is a 3×3 unit matrix. In the program DYANA, ρ =5Å and $m_k = 10\sqrt{n_k}m_0$, in which n_k is the number of atoms in the rigid body k, and $m_0 = 1.66 \times 10^{-27}$kg.

The equations of motion for a classical mechanical system with generalized coordinates are the Lagrange equations [7]:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}_k}\right) - \frac{\partial L}{\partial \theta_k} = 0 \quad (k = 1, 2, ..., n), \tag{2.18}$$

in which $L = E_{kin} - E_{pot}$. Therefore, the equation for $n \times n$ mass matrix $M(\theta)$ and n-dimensional vector $C(\theta, \dot{\theta})$ can be derived:

$$M(\theta)\ddot{\theta} + C(\theta, \dot{\theta}) = 0. \tag{2.19}$$

In order to obtain the torsional accelerations $\ddot{\theta}$, we need to integrate Equations (2.18) and (2.19) at each time step, which has extremely high computational complexity in large molecular systems. Therefore, in the program DYANA, the recursive algorithm in [34] has been used to calculate the torsional accelerations.

The algorithm [34] calculates three six-dimensional vectors a_k, e_k, z_k and two 6×6 matrices P_k and ϕ_k with $(k = 1, 2, ..., n)$ for the initialization process:

$$a_k = \begin{bmatrix} (\vec{w}_k \wedge \vec{e}_k)\dot{\theta}_k \\ \vec{w}_{p(k)} \wedge (\vec{v}_k - \vec{v}_{p(k)}) \end{bmatrix}, \tag{2.20}$$

$$e_k = \begin{bmatrix} \vec{e}_k \\ \vec{0} \end{bmatrix},$$

$$z_k = \begin{bmatrix} \vec{w}_k \wedge \mathbf{I}_k\vec{w}_k \\ (\vec{w}_k \cdot m_k\vec{Y}_k)\vec{w}_k - w_k^2 m_k\vec{Y}_k \end{bmatrix},$$

$$P_k = \begin{bmatrix} \mathbf{I}_k & m_k\mathbf{A}(\vec{Y}_k) \\ -m_k\mathbf{A}(\vec{Y}_k) & m_k\mathbf{1}_3 \end{bmatrix},$$

$$\phi_k = \begin{bmatrix} \mathbf{1}_3 & \mathbf{A}(\vec{r}_k - \vec{r}_{p(k)}) \\ \mathbf{0}_3 & \mathbf{1}_3 \end{bmatrix},$$

where $\mathbf{0}_3$ is a 3×3 zero matrix, and $\mathbf{A}(x)$ is an antisymmetric 3×3 matrix with the property that for all vectors $\vec{y}$: $\mathbf{A}(\vec{x})\vec{y} = \vec{x} \wedge \vec{y}$.

Next, the following equations are calculated recursively over all the rigid bodies in the backward direction, $k = n, n-1, ..., 1$:

$$\begin{aligned} D_k &= e_k \cdot P_k e_k, \\ G_k &= P_k e_k / D_k, \\ \varepsilon_k &= -e_k \cdot (z_k + P_k a_k) - \frac{\partial V}{\partial \theta_k}, \\ P_{p(k)} &\longleftarrow P_{p(k)} + \phi_k (P_k - G_k e_k^T P_k) \phi_k^T, \\ z_{p(k)} &\longleftarrow z_{p(k)} + \phi_k (z_k + P_k a_k + G_k \varepsilon_k), \end{aligned} \tag{2.21}$$

in which D_k and ε_k are scalars, G_k is a six-dimensional vector, and arrow $\longleftarrow$ is to assign the right-hand side to the left-hand side.

Finally, a recursive loop over all the rigid bodies in the forward direction will be made to derive torsional accelerations, $k = 1, 2, ..., n$:

$$a_k = \phi_k^T a_{p(k)},$$

$$\ddot{\theta}_k = \varepsilon_k / D_k - G_k \cdot a_k, \tag{2.22}$$

$$a_k \longleftarrow a_k + e_k \ddot{\theta}_k + a_k. \tag{2.23}$$

The auxiliary quantities D_k, G_k and ε_k are calculated from Equation (2.21). Acceleration is given by Equation (2.22). In this loop, auxiliary quantity a_k is updated by Equation (2.23), with initial value a_0 being equal to the zero vector.

The recursive algorithm used in DYANA follows these steps:

(1) According to the torsion angles $\theta(t)$ at time t, calculate the Cartesian coordinates of all atoms [24].

(2) Calculate the potential energy function $E_{pot}(t) = E_{pot}(\theta(t))$, and its gradient $\nabla E_{pot}(t)$.

(3) Get current time step, $\Delta t = \lambda_\varepsilon \Delta t'$, in the time sequence $(t - \Delta t') \rightarrow t \rightarrow (t + \Delta t)$, $\Delta t'$ denotes the previous time step of Δt, and scaling factor

$$\lambda_\varepsilon = \min\left(\lambda_\varepsilon^{\max}, \sqrt{1 + \frac{\varepsilon^{\text{ref}} - \varepsilon(t)}{\tau \varepsilon(t)}}\right),$$

where $\lambda_\varepsilon^{\max} = 1.025$ is the upper bound of the scaling factor λ_ε, the time constant $\tau = 20$, ε^{ref} denotes the reference value for the relative accuracy of

energy conservation, and $\varepsilon(t)$ as the relative change of the total energy in the time step $\Delta t'$ is given by

$$\varepsilon(t) = \left| \frac{E(t) - E(t - \Delta t')}{E(t)} \right|.$$

(4) Control the temperature by scaling the torsional velocities,

$$\begin{aligned} \dot{\theta}(t - \Delta t'/2) &\leftarrow \lambda_T \dot{\theta}(t - \Delta t'/2), \\ \dot{\theta}(t) &\leftarrow \lambda_T \dot{\theta}(t). \end{aligned}$$

The scaling factor λ_T is calculated as

$$\lambda_T = \sqrt{1 + \frac{T^{\text{ref}} - T(t)}{\tau T(t)}},$$

and the current temperature $T(t)$ is given by

$$T(t) = \frac{2E_{kin}(t)}{nk_B},$$

where Boltzmann constant $k_B = 1.38 \times 10^{-23}\text{JK}^{-1}$, and the kinetic energy is calculated by $E_{kin}(t) = E_{kin}(\theta(t), \dot{\theta}_e(t))$. Temperature and time-step control can be turned off by setting τ to ∞.

(5) Determine the torsional accelerations $\ddot{\theta}(t)$ by Equations (2.20–2.22).

(6) Use a leap-frog scheme to calculate new angular velocities at half time-step:

$$\dot{\theta}(t + \Delta t/2) = \dot{\theta}(t - \Delta t'/2) + \frac{\Delta t + \Delta t'}{2} \ddot{\theta}(t),$$

and to calculate the new estimated angular velocities at full time-step:

$$\dot{\theta}_e(t + \Delta t) = \left(1 + \frac{\Delta t}{\Delta t + \Delta t'}\right) \dot{\theta}(t + \Delta t/2) - \frac{\Delta t}{\Delta t + \Delta t'} \dot{\theta}(t - \Delta t'/2).$$

(7) Calculate new torsion angles:

$$\theta(t + \Delta t) = \theta(t) + \Delta t \dot{\theta}(t + \Delta t/2).$$

The algorithm will loop recursively over the above seven steps. Time step value will be updated after execution of Step 7. The initial values of variables are: $t = 0$, $\Delta t' = \Delta t$, and $\dot{\theta}_e(0) = \dot{\theta}(-\Delta t/2)$. Initial value for the torsional

Table 2.6: Computation time (in seconds) for DYANA structure calculations of the proteins BPTI and cyclophilin A on different computers [25, 27].

Computer	BPTI	Cyclophilin A
NEC SX-4	13	36
DEC Alpha 8400 5/300	20	86
SGI Indigo2 R10000 (175 MHz)	23	127
IBM RS/6000-590	35	141
Cray J-90	44	141
Convex Exemplar	44	177
Hewlett-Packard 735	47	209

velocities $\dot{\theta}(-\Delta t/2)$ are determined by a normal distribution with zero mean value and a standard deviation that guarantees the initial temperature has a initial value $T(0)$.

With regard to the efficiency of structure calculation by program DYANA, the computation time is listed in Table 2.6 [25, 27]. The computation time is below 1 mine for small proteins such as BPTI, and only 3.5 minutes for proteins such as cyclophilin A, on the generally available computers listed in Table 2.6. Since an NMR structure calculation always results in a group of comformers [50], it is more efficient to run the calculations in parallel. In DYANA, parallel computing is implemented using language INCLAN. Using a dedicated Cray J-90 computer with eight processors, the elapsed time for the calculation of 100 BPTI conformers on n=1, ..., 8 processors was closely proportional to $1/n$, which is the theoretically achievable optimum [27].

Chapter 3

Experimental Results

In this chapter, some experimental results are presented that follow the steps introduced in Chapter 2. Each step will be introduced in detail from the experimental aspect, including I/O files and experimental procedure. Experimental results given here are from experiments on the protein `MT0776` conducted by Wishart Research Group [5]. `MT0776` is a protein from an organism called "Methanobacterium thermoautotrophicum", which is a methane producing organism from the Archaea kingdom of bacteria. It is the 776th protein found on the genome sequence, so it was designated `MT0776` [3].

3.1 Data Processing

The input data for the entire experiment are the signals recorded, digitized, and amplified by an NMR spectrometer. The signals are usually called Free Induction Decay (FID), which are time-domain spectra as illustrated in Figure 3.1. The data acquisition software for `MT0776` experiments are VNMR.

Data processing is utilized to perform the conversion between time-domain and frequency-domain spectra, noise filtering, resolution increase, and line-shape smoothing. The functionalities of the existing data processing software include Fourier transform, apodization, linear prediction and phase correction for 1D-3D spectra. The data processing could also be performed by a pipeline of unix command, as in the program NMRPipe, where each unix command

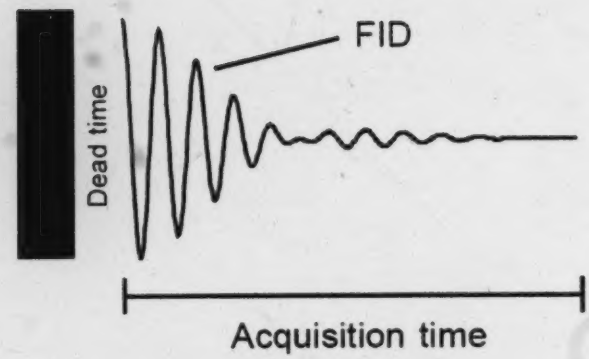

Figure 3.1: One-dimensional Time-domain Spectrum FID

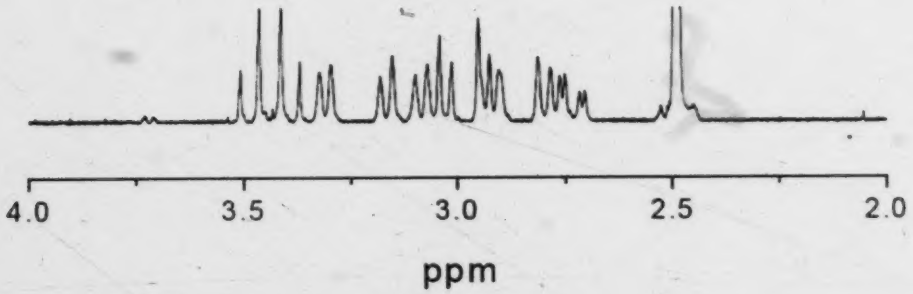

Figure 3.2: One-dimensional Frequency-domain Spectrum

represents a single data processing step. The data processing software using in the experiments of MT0776 is NMRPipe. The output files from data processing are 1D-3D frequency-domain spectra. Figure 3.2 is an example of one-dimensional frequency-domain spectrum.

3.2 Peak Picking

The peak picking program reads in 2D or 3D spectra data and identifies potential peaks, along with their intensities. For protein MT0776, two 3D spectra CBCAcoNH and HNCACB, and one 2D spectra 15N-HSQC (heteronuclear single quantum correlation) are used for backbone assignment. The other two 3D spectra CcoNH and HccoNH are used for side-chain assignment. In 2D

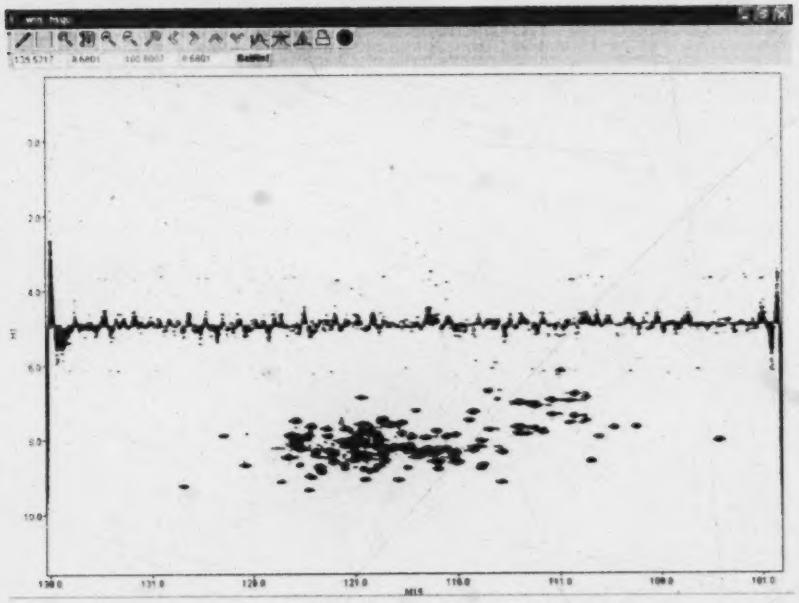

Figure 3.3: Two-dimensional HSQC (heteronuclear single quantum correlation) Spectrum before Peak Picking

15N-HSQC (Figure 3.3), each peak contains two entries, one for amide proton (HN) chemical shift, the other for the directly attached nitrogen (NH) chemical shift. In 3D CBCAcoNH, each peak contains three entries, for HN, NH, and carbon chemical shift, respectively. The entry for carbon includes carbon alpha (CA) and carbon beta (CB) from the same amino acid residue. Those three entries are the same as in the 3D HNCACB; the difference is the carbon entry, which includes CA and CB from two adjacent amino acid residues in 3D HNCACB. In 3D spectra CcoNH, three entries of each peak correspond to HN, NH, and side-chain carbon chemical shift, respectively. However, in 3D spectra HccoNH, the third entry of each peak is a side-chain hydrogen chemical shift. Once all the spectra needed to build the protein structure are ready, we can proceed to the peak picking process. Other complementary spectra are 2D 13C-HSQC, 3D HNCA, 3D HNCO, and 3D HNHA. 3D HNCA and 3D HNCO have the same entries as 3D HNCACB. In 3D HNHA, the three entries are (N, HN, HA); the entries for 2D 13C-HSQC are (C, H).

The first stage in the peak picking process is to filter the noise based on the possible chemical shift ranges for each spectrum. The possible peak should be found in the range of [6.5ppm–12ppm] for HN, [40ppm–65ppm] for CA, [60ppm–75ppm] for CB, [12ppm–40ppm] for side-chain carbon, and [90ppm–137ppm] for backbone N. Therefore, peak picking should be restrained to these ranges. Figure 3.4 is an example of resulting spectrum.

The next stage is the automated peak picking process which is usually performed by software alone. This process usually identifies the coordinates and intensities of the potential peaks, and does not have a high accuracy. This means that the peak list generated by software cannot be used directly for peak assignment, and the last stage peak list correction is performed to ensure a qualified peak list. For the HNCACB spectrum of MT0776, the peak list resulting from the automated peak picking process includes 2944 peaks with intensity threshold value at 1000000 on the NMRDraw standard.

In the final stage, knowledge of correlations between different spectra, dipolar coupling, and scalar coupling are used to correct the peak list following the automated peak picking process. In spectrum HNCACB, the CA and CB peaks from the i-th and $(i-1)$-th residue should have corresponding peaks on two different planes from spectrum CBCAcoNH, and the peaks from spectrum CBCAcoNH could be used to verify potential peaks in spectrum HNCACB. The CB peaks appearing in spectrum CBCAcoNH should have negative intensities, shown in red color in spectra. In spectrum HNCACB, the peaks from the residue i are usually more intense than those from the residue $(i-1)$. Making a qualified peak list is one of the most crucial stages in the whole procedure. The resulting peak list from this stage includes only 2730 real peaks out of the original 2944 peaks in the HNCACB spectrum of protein MT0776 data. These are the result coming from peak picking software NMRView.

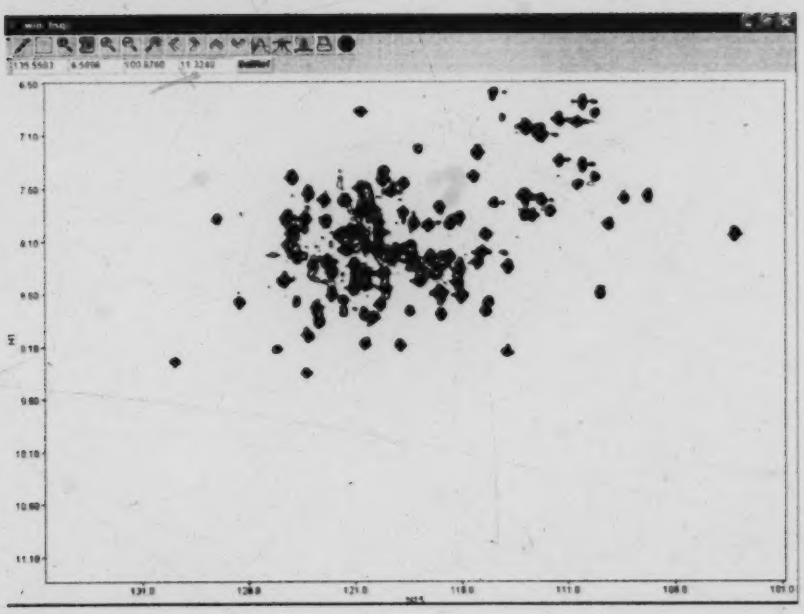

Figure 3.4: Two-dimensional HSQC (heteronuclear single quantum correlation) Spectrum after Peak Picking

3.3 Peak Assignment

The peak assignment process is use to assign peaks to amino acid residues on polypeptide chains of a protein. The first stage of the assignment is *peak grouping*, which groups the chemical shifts from the same amino acid residue to a spin system. The input file will be the peak lists generated by the peak picking process. Peak grouping could therefore be done by searching for identical backbone NH and HN chemical shift values among the peak lists of HSQC, CBCAcoNH, and HNCACB. The backbone NH and HN chemical shifts are collected from the 2D HSQC spectrum. Backbone CA and CB chemical shifts are obtained from HNCACB with complementary information from CBCAcoNH.

Once the spin system is formed, the connection between spin systems from adjacent amino acid residues will be set up by the connectivity information of CB and CA peaks in the second stage. CB_i, CA_i, CB_{i-1}, and CA_{i-1} are

identified from spectra HNCACB and CBCAcoNH. An output file will be written out, which includes information about the two adjacent spin systems.

The third stage of the assignment is to generate possible spin system chains in order to map to the polypeptide chain. The possible amino acid type of each spin system is determined based on the knowledge of chemical shift statistics (Table 3.1) and spin topology. With the connectivity information derived from the preceding stage, possible spin system chains are generated.

The last stage is to map those possible spin system chains to the protein sequence by a scoring scheme. The scoring scheme used in this experiment takes four factors into account. They are the length of the spin system chain, deviations of the chemical shifts on chemical shift statistics, intensities of the peaks for each spin system, and the similarity to the sequence from a homology protein, if available. The mapping process starts from the spin systems which can be unambiguously assigned, as illustrated in Table 3.2. Scores have been assigned by the scoring scheme to each mapping between spin system chains and segments on the protein sequence. Finally, the greedy search is applied to complete the assignment process by searching from the highest score. Once the backbone assignment (i.e., sequential assignment) is complete, side-chain carbons' and hydrogens' chemical shifts are assigned from spectra CcoNH and HccoNH, respectively. The peak assignment based on the HNCACB spectrum of `MT0776` results in 8 unassigned amino acid residues whose chemical shifts are not found at all. These unassigned amino acid residues will be made up by the peak assignment resulting from the CBCAcoNH spectrum. Appendix A is an example of the output peak list file from peak assignment on 15N-HSQC spectrum.

3.4 NOESY Assignment

Nuclear Overhauser Effect (NOE) is a common phenomenon for pairs of nuclei of any type with spatial distance between them shorter than 5Å. Nuclear overhauser effect between pairs of hydrogen atoms is the easiest one to detect

Table 3.1: Amino acid chemical shift statistics I [1]

Amino Acid	Atom Name	Atom Type	Number of Shifts	Minimum Shift	Maximum Shift	Average Shift	Standard Deviation
ALA	H	H	14184	0.95	12.25	8.19	0.63
ALA	HA	H	12028	1.24	8.62	4.26	0.45
ALA	HB	H	11081	-2.29	3.70	1.37	0.28
ALA	C	C	6966	49.30	185.42	177.76	2.69
ALA	CA	C	10362	42.82	99.00	53.20	2.13
ALA	CB	C	9465	0.00	99.00	19.04	2.54
ALA	N	N	11688	18.28	219.60	123.24	4.56

Table 3.2: Amino acid chemical shift statistics II [5]

	CA chemical shift (ppm)	CB chemical shift (ppm)
Glycines	43-48	N/A
Serine	57-61	61-65
Threonine	60-68	63-70

in NMR experiments. Therefore, both entries of NOE spectrum are chemical shift values of hydrogen atoms. As the distance constraints source, NOESY experiments are the most important experiments for the structure calculation. However, NOESY peak assignment will not achieve good results without the assistance of the sequential assignment results. NOESY peak picking and assignment are the same processes, as the sequential assignment described in the previous two sections. Based on the dipolar coupling knowledge, peaks in NOESY are predictable with known chemical shifts on the polypeptide chain of the protein. For instance, if a peak has corresponding hydrogen chemical shifts to the hydrogen atoms on the protein sequence for both entries of the peak, then that's a real peak. Therefore, in the experiments, artificial peaks could be partially filtered in this way. Three spectra 2D 1H-1H NOESY, 3D 1H-15N NOESY and 3D 13C-NOESY are used for NOESY assignment. The entries of peaks from the spectrum 2D 1H-1H NOESY, 3D 1H-15N NOESY, and 3D 13C-NOESY are (H, H),(N, H, H) and (C, H, H), respectively.

The peak list output from the NOESY assignment includes information about the peak positions, hydrogen chemical shift values, and peak intensities. Using the NOESY analysis tool, parameters for the cut-off intensities for each classes of NOE peaks could be adjusted by the user, and the NOE constraints file can be written out in different formats as shown in Appendix B. For instance, the constraint in the NOE distance constraint file are the distance bounds between two atoms {(resid 2 and name HN) (resid 2 and name HA)}. The lower bound is (3.0-1.2), and the upper bound is (3.0+1.5).

3.5 3D Structure Calculation

The algorithm used for structure calculation is molecular dynamics in Cartesian space. The representative software using here is XPLOR. The first stage of structure calculation is to generate a homology model (Swiss PDB model [22]) of the protein. Protein Structure Files (PSF) are generated as a summary of the atom type, mass, partial charge, and connectivity of the molecular

system. PSF files are generated from the original PDB file, using SwissPDB in combination with the topology file. The topology file is the file which shows the skeleton of the protein; it is usually generated from a group of homology protein topologies. A template coordinate set is also generated with the PSF file as input; it produces an arbitrary extended conformation with ideal geometry, and writes out as a template file.

The second stage is the simulated annealing procedure, which simulates the slowly cooling procedure from a very high temperature. The input files include the structure file PSF, the template file, the NOE distance constraints file and the dihedral angle constraints file (Appendix C). For example, four atoms {(resid 1 and name c) (resid 2 and name n) (resid 2 and name ca) (resid 2 and name c)} for each constraint in the file are four atoms involved in this dihedral angle on the sequential order of protein sequence. The numbers in the dihedral angle constraints file {1.0, -120.0, 10.0, 2} represent the energy constant k_a in Equation 2.11, the value of this dihedral angle, the tolerance of this angle, and the exponent of the dihedral angle term in the potential energy function Equation 2.11 respectively. The parameters involving starting temperature, time step, the number of structures generated, and other parameters related to the energy terms could be adjusted by the user. The output of this stage is a group of conformations of the atoms; another simulated annealing protocol will be executed to refine the resulting conformations. The output PDB files include information representing the overall energy value, energy value for each energy term, constraint violations, and many parameters describing the protein structure. Figure 3.5 shows several conformers of protein `MT0776`. There are slight differences between conformers in the same column, but big differences between two columns. Therefore, structure calculation needs to run a large number of times to get a group of good conformers.

The last stage will be the checking procedure which checks the acceptable protein structures. In practice, it is impossible to determine which conformation represents the exact structure of the protein. By using a protein checking program, we will be able to distinguish which conformation is the closest to the real structure by different sets of standards.

Figure 3.5: Several conformers of protein MT0776

Chapter 4

Conclusions and Future Work

This chapter presents conclusions and possible future work.

4.1 Conclusions

Nuclear Magnetic Resonance (NMR) spectroscopy is one of a few available techniques which can be used to determine three-dimensional structures of macromolecules. Although NMR spectroscopy may not achieve the same accuracy as X-ray crystallography, with the acceleration by computational and molecular modelling methods, NMR spectroscopy competes X-ray crystallography for becoming the dominant high-throughput technique for protein structure determination in the future.

This dissertation provided a global picture of the procedure of protein structure determination using Nuclear Magnetic Resonance spectroscopy from computer science point of view. The procedure of protein structure determination based on NMR spectroscopy experiments includes data acquisition, data processing, peak picking, peak assignment, and structure determination. Besides the detailed process of each stage of the procedure, several available programs and software were discussed and compared with each other in this dissertation. Some innovative progresses on the peak assignment process made in our group were presented further. The complete procedure was demonstrated in Chapter 3, using the protein `MT0776` (from Wishart Research Group [5]) as an

example.

In conclusion, NMR spectroscopy is an extremely powerful tool for solving macromolecular structures. Innovative computational and molecular modelling methods dramatically decreased the time of resolving protein structure by NMR from years to weeks. Whereas, NMR spectroscopy is a semi-automated process. There are no completely mature or fully automated programs or software for the entire structure determination procedure. In order to generate an acceptable protein structure, considerable manual work is involved in the procedure. As technique develops, improved computational methods for each process can be expected.

4.2 Future Work

Determining protein structure based on NMR spectroscopy is a rewarding and challenging research problem. Several projects can serve as follow up work: (1) combination of the existing software or programs; (2) automation of each process in NMR spectroscopy; (3) complete pipeline of NMR spectroscopy on protein structure determination.

First of all, we may be able to combine different software or programs together to produce a better method. For the structure determination process, distance geometry could be used to generate starting structures for torsion angle dynamics algorithm. Although torsion angle dynamics algorithm is the most widely used structure calculation algorithm, it is not able to deal with the structure calculation of the closed flexible rings part of the protein. The combination of these three algorithms can lead to a better solution for the 3-D structure calculation process.

Second aspect of the future work can be the automation of each process in NMR spectroscopy. Considerable manual work is involved in each process more or less. Therefore, automation of the procedure is fairly crucial for the future development of the NMR spectroscopy.

In addition, the development of a complete pipeline of protein structure

determination by NMR spectroscopy is also a potential project. There are a number of software and programs available for each process of NMR spectroscopy for protein structure determination. However, input and output files from various software are not compatible with each other. It takes a long period of time to learn how to use various software. Once a complete pipeline is generated, the procedure could be highly accelerated.

Bibliography

[1] BMRB Amino Acid Chemical Shift Statistics. In "http://www.bmrb.wisc.edu/",.

[2] MathWorld. In "http://mathworld.wolfram.com/",.

[3] NiceProt View of TrEMBL: O26870. In "http://ca.expasy.org/cgi--bin/niceprot.pl?O26870",.

[4] Wikipedia, the free encyclopedia. In "http://en.wikipedia.org/wiki/Main_Page",.

[5] Wishart research group. In "http://redpoll.pharmacy.ualberta.ca/",.

[6] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering Points To Identify the Clustering Structure. In *Proceedings of ACM SIGMOD'99 International Conference on Management of Data*, pages 49–60, 1999.

[7] V. I. Arnold. *Mathematical methods of classical mechanics*. Springer, 1978.

[8] C. Bartels, M. Billeter, P. Guntert, and K. Wuthrich. Automated Sequence–specific Assignment of Homologous Proteins using the program GARANT. *Journal of Biomolecular NMR*, 7:207–213, 1996.

[9] C. Bartels, P. Guntert, M. Billeter, and K. Wuthrich. GARANT–a General Algorithm for Resonance Assignment of Multidimensional Nuclear Magnetic Resonance Spectra. *Journal of Computational Chemistry*, 18:139–149, 1997.

[10] M. Billeter, T. Schaumann, W. Braun, and K. Wuthrich. Restrained energy refinement with two different algorithms and force fields of the structure of the α–amylase inhibitor tendamistat determined by NMR in solution. *Biopolymers*, 29:695–706, 1990.

[11] W. Braun, C. Bosch, L. R. Brown, N. Go, and K. Wuthrich. Combined use of proton–proton overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. *Biochimica et Biophysics Acta*, 667:377–396, 1981.

[12] W. Braun and N. Go. Calculation of protein conformations by proton–proton distance constraints. *Journal of Molecular Biology*, 186:611–626, 1985.

[13] A. T. Brunger. *X–PLOR, Version 3.1. A system for X–ray Crystallography and NMR*. Yale University Press, 1992.

[14] J. Cavanagh, W. Fairbrother, A. Palmer III, and N. Skelton. *Protein NMR Spectroscopy: Principles & Practice*. Academic Press, 1995.

[15] B. E. Coggins and P. Zhou. PACES: Protein sequential assignment by computer-assisted exhaustive search. *Journal of Biomolecular NMR*, 26:93–111, 2003.

[16] F. Cordier and S. Grzesiek. Direct observation of hydrogen bonds in proteins by interresidue (3h)J(NC') scalar couplings. *Journal of the American Chemical Society*, 121:1601–1602, 1999.

[17] F. Cordier, M. Rogowski, S. Grzesiek, and A. Bax. Observation of through-hydrogen-bond (2h)J(HC') in a perdeuterated protein. *Journal of Magnetic Resonance*, 140:510–512, 1999.

[18] G. Cornilescu, F. Delaglio, and A. Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *Journal of Biomolecular NMR*, 13:289–302, 1999.

[19] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[20] A. J. Dingley and S. Grzesiek. Direct observation of hydrogen bonds in nucleic acid base pairs by internucleotide (2)J(NN) couplings. *Journal of the American Chemical Society*, 120:8293–8297, 1998.

[21] A. J. Fischman, D. H. Live, H. R. Wyssbrod, W. C. Agosta, and D. Cowburn. Torsion angles in the cystine bridge of oxytocin in aqueous solution. Measurements of circumjacent vicinal couplings between ^{1}H, ^{13}C, and ^{15}N. *Journal of the American Chemical Society*, 102:2533–2539, 1980.

[22] N. Guex and M. C. Peitsch. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, 18:2714–2723, 1997.

[23] W. F. Van Gunsteren and H. J. C. Berendsen. Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry. *Angewandte Chemie International edition*, 29:992–1023, 1990.

[24] P. Guntert. *Neue Rechenverfahren fur die Proteinstrukturbestimmung mit Hilfe der magnetischen Kernspinresonanz*. PhD thesis, 10135 ETH Zurich, 1993.

[25] P. Guntert. Structure calculation of biological macromolecules from NMR data. *Quarterly Reviews of Biophysics*, 31:145–237, 1998.

[26] P. Guntert, W. Braun, and K. Wuthrich. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *Journal of Molecular Biology*, 217:517–530, 1991.

[27] P. Guntert, C. Mumenthaler, and K. Wuthrich. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *Journal of Molecular Biology*, 273:283–298, 1997.

[28] P. Guntert, Y. Q. Qian, G. Otting, M. Muller, W. Gehring, and K. Wuthrich. Structure determination of the Antp ($C_{39} \rightarrow S$) homeodomain from nuclear magnetic resonance data in solution using a novel strategy for the structure calculation with the programs diana, caliba, habas and glomsa. *Journal of Molecular Biology*, 217:531–540, 1991.

[29] P. Guntert, M. Salzmann, D. Braun, and K. Wuthrich. Sequence–specific NMR Assignment of proteins by global fragment mapping with the program MAPPER. *Journal of Biomolecular NMR*, 18:129–137, 2000.

[30] T. F. Havel and K. Wuthrich. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular $^1H - ^1H$ proximities in solution. *Bulletin of Mathematical Biology*, 46:673–698, 1984.

[31] T. Herrmann, P. Guntert, and K. Wuthrich. Protein NMR structure determination with automated NOE–identification in the NOESY spectra using the new software ATNOS. *Journal of Biomolecular NMR*, 24:171–189, 2002.

[32] T. Herrmann, P. Guntert, and K. Wuthrich. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of Molecular Biology*, 319:209–227, 2002.

[33] J S. Hu and A. Bax. Determination of ϕ and χ^1 angles in proteins from $^{13}C - ^{13}C$ three–bond J couplings measured by three–dimensional heteronuclear NMR. How planar is the peptide bond? *Journal of the American Chemical Society*, 119:6360–6368, 1997.

[34] A. Jain, N. Vaidehi, and G. Rodriguez. A fast recursive algorithm for molecular dynamics simulation. *Journal of Computational Physics*, 29:992–1023, 1990.

[35] R. Koradi, M. Billeter, M. Engeli, P. Guntert, and K. Wuthrich. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *Journal of Magnetic Resonance*, 135:288–297, 1998.

[36] R. Koradi, M. Billeter, and K. Wuthrich. MOLMOL: a program for display and analysis of macromolecular structures. *Journal of Molecular Graph*, 14:51–55, 1996.

[37] M. Leutner, R. M. Gschwind, J. Liermann, C. Schwarz, G. Gemmecker, and H. Kessler. Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *Journal of Biomolecular NMR*, 11:31–43, 1998.

[38] A. De Marco, M. Llinas, and K. Wuthrich. $^1H - ^{15}N$ spin–spin couplings in alumichrome. *Biopolymers*, 17:2727–2742, 1978.

[39] A. De Marco, M. Llinas, and K. Wuthrich. Analysis of the ^1H–NMR spectra of ferrichrome peptides. I. The non–amide protons. *Biopolymers*, 17:617–636, 1978.

[40] J. L. Markley, A. Bax, Y. Arata, C. W. Hilbers, R. Kaptein, B. D. Sykes, P. E. Wright, and K. Wuthrich. Recommendations for the presentation of NMR structures of proteins and nucleic acids. *Journal of Molecular Biology*, 280:933–952, 1998.

[41] M. Ottiger, T. Szyperski, L. Luginbuhl, C. Ortenzi, P. Luporini, R. A. Bradshaw, and K. Wuthrich. The NMR solution structure of the pheromone Er–2 from the ciliated protozoan Euplotes raikovi. *Protein Science*, 3:1515–1526, 1994.

[42] N. Tjandra, J. G. Omichinski, A. M. Gronenborn, G. M. Clore, and A. Bax. Use of dipolar $^{1}H - ^{15}N$ and $^{1}H - ^{13}C$ couplings in the structure determination of magnetically oriented macromolecules in solution. *Nature Structural Biology*, 4:732–738, 1997.

[43] X. Wan and G. H. Lin. CISA: Combined NMR Resonance Connectivity Information Determination and Sequential Assignment. *Journal of Biomolecular NMR*, submitted in 2005.

[44] X. Wan, T. Tegos, and G. H. Lin. Histogram–based scoring schemes for protein NMR resonance assignment. *Journal of Bioinformatics and Computational Biology*, 2:747–764, 2004.

[45] A. C. Wang and A. Bax. Reparametrization of the Karplus relation for $^{3}J(H^{\alpha} - N)$ in peptides from uniformly $^{13}C/^{15}N$ enriched human ubiquitin. *Journal of the American Chemical Society*, 117:1810–1813, 1995.

[46] A. C. Wang and A. Bax. Determination of the backbone dihedral angles ϕ in human ubiquitin from reparametrized empirical Karplus equations. *Journal of the American Chemical Society*, 118:2483–2494, 1996.

[47] M. P. Williamson, T. F. Havel, and K. Wuthrich. Solution conformation of proteinase inhibitor IIa from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry. *Journal of Molecular Biology*, 182:295–315, 1985.

[48] D. S. Wishart and B. D. Sykes. The 13C Chemical–Shift Index: A simple method for the identification of protein secondary structure using 13C chemical–shift data. *Journal of Biomolecular NMR*, 4:171–180, 1994.

[49] D. S. Wishart, B. D. Sykes, and F. M. Richards. The Chemical Shift Index: A Fast and Simple Method of the Assignment of Protein Secondary Structure through NMR spectroscopy. *Biochemistry*, 31:1647–1651, 1992.

[50] K. Wuthrich. *NMR of Proteins and Nucleic Acids*. Wiley, 1986.

[51] K. Wuthrich, M. Billeter, and W. Braun. Polypeptide Secondary Structure Determination by Nuclear Magnetic Resonance Observation of Short Proton–Proton Distances. *Journal of Molecular Biology*, 180:715–740, 1984.

[52] Y. Xu, D. Xu, D. Kim, V. Olman, J. Razumovskaya, and T. Jiang. Automated assignment of backbone NMR peaks using constrained bipartite matching. *IEEE Computing in Science & Engineering*, 4:50–62, 2002.

[53] D. E. Zimmerman, C. A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. Shimotakahara, C. Chien, R. Powers, and G. T. Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Biomolecular NMR*, 26:93–111, 2003.

Appendix A

Peak List Sample

```
label dataset sw sf
N15 H1
hsqc15.nv
{1800.00} {3000.30}
{60.7970} {599.9330}
N15.L N15.P N15.W N15.B N15.E N15.J N15.U H1.L H1.P H1.W H1.B H1.E H1.J
H1.U vol int stat comment flag0
0 {71.n} 123.299 0.183 0.183 ++ 0.000 {?} {71.hn} 9.354 0.044 0.044 ++
0.000 {?} 0.08073 0.08073 0 {?} 0
1 {78.N} 113.842 0.307 0.318 ++ 0.000 {?} {78.HN} 9.153 0.042 0.043 ++
0.000 {?} 0.11845 0.11845 0 {?} 0
2 {54.N} 118.917 0.242 0.464 ++ 0.000 {?} {54.HN} 9.085 0.036 0.061 ++
0.000 {?} 0.24631 0.24631 0 {?} 0
3 {9.N} 120.542 0.175 0.417 ++ 0.000 {?} {9.HN} 9.070 0.030 0.065 ++ 0.000
{?} 0.35988 0.35988 0 {?} 0
4 {12.N} 123.228 0.214 0.333 ++ 0.000 {?} {12.HN} 8.999 0.052 0.072 ++
0.000 {?} 0.17864 0.17864 0 {?} 0
5 {35.N} 122.705 0.318 0.752 ?  0.000 {?} {35.HN} 8.861 0.026 0.044 ++
0.000 {?} 0.24996 0.24996 0 {?} 0
6 {56.N} 121.489 0.268 0.491 ++ 0.000 {?} {56.HN} 8.769 0.041 0.057 ++
0.000 {?} 0.20999 0.20999 0 {?} 0
7 {88.N} 120.203 0.202 0.636 ?  0.000 {?} {88.HN} 8.832 0.033 0.075 ++
0.000 {?} 0.45919 0.45919 0 {?} 0
8 {72.n} 116.994 0.259 0.633 ++ 0.000 {?} {72.hn} 8.792 0.034 0.054 ++
0.000 {?} 0.24283 0.24283 0 {?} 0
9 {58.N} 120.585 0.297 0.323 ++ 0.000 {?} {58.HN} 8.784 0.050 0.054 ++
0.000 {?} 0.12572 0.12572 0 {?} 0
10 {28.N} 114.902 0.178 0.514 ++ 0.000 {?} {28.HN} 8.763 0.037 0.075 ++
0.000 {?} 0.43272 0.43272 0 {?} 0
11 {6.N} 122.860 0.279 0.436 ++ 0.000 {?} {6.HN} 8.733 0.033 0.050 ++ 0.000
{?} 0.19893 0.19893 0 {?} 0
12 {99.N} 119.821 0.220 0.520 ?  0.000 {?} {99.HN} 8.701 0.034 0.071 ++
0.000 {?} 0.40770 0.40770 0 {?} 0
......
```

Appendix B

NOE Distance Constraints File Sample

```
set message=on echo=on end
!Intra-residue noes
assign (resid 2 and name HN )(resid 2 and name HA ) 3.0 1.2 1.5
assign (resid 2 and name HN )(resid 2 and name HB# ) 4.0 2.2 1.0
assign (resid 2 and name HG# )(resid 2 and name HB# ) 4.0 2.2 1.0
assign (resid 3 and name HN )(resid 3 and name HA ) 4.0 2.2 1.0
assign (resid 4 and name HN )(resid 4 and name HA ) 3.0 1.2 1.5
assign (resid 4 and name HN )(resid 4 and name HB# ) 4.0 2.2 1.5
assign (resid 4 and name HD# )(resid 4 and name HB# ) 4.0 2.2 1.5
assign (resid 4 and name HN )(resid 4 and name HD# ) 4.0 2.2 2.5
assign (resid 5 and name HN )(resid 5 and name HB# ) 4.0 2.2 1.5
assign (resid 6 and name HN )(resid 6 and name HA ) 4.0 2.2 1.0
assign (resid 7 and name HN )(resid 7 and name HA ) 4.0 2.2 1.0
assign (resid 7 and name HN )(resid 7 and name HB# ) 4.0 2.2 1.0
assign (resid 8 and name HN )(resid 8 and name HA ) 4.0 2.2 1.0
assign (resid 9 and name HN )(resid 9 and name HA ) 4.0 2.2 1.0
assign (resid 9 and name HN )(resid 9 and name HB# ) 4.0 2.2 1.0
assign (resid 11 and name HN )(resid 11 and name HB# ) 4.0 2.2 1.5
assign (resid 12 and name HN )(resid 12 and name HB# ) 4.0 2.2 1.5
assign (resid 12 and name HN )(resid 12 and name HA ) 4.0 2.2 1.0
assign (resid 13 and name HN )(resid 13 and name HA ) 4.0 2.2 1.0
assign (resid 13 and name HN )(resid 13 and name HB# ) 4.0 2.2 1.5
assign (resid 14 and name HN )(resid 14 and name HA# ) 4.0 2.2 1.5
assign (resid 15 and name HN )(resid 15 and name HA ) 4.0 2.2 1.5
assign (resid 16 and name HN )(resid 16 and name HA ) 4.0 2.2 1.5
assign (resid 16 and name HN )(resid 16 and name HD# ) 4.0 2.2 1.0
assign (resid 16 and name HD# )(resid 16 and name HE# ) 4.0 2.2 1.5
assign (resid 17 and name HN )(resid 17 and name HB# ) 4.0 2.2 1.5
assign (resid 18 and name HN )(resid 18 and name HA ) 4.0 2.2 1.0
assign (resid 19 and name HN )(resid 19 and name HA ) 4.0 2.2 1.0
assign (resid 22 and name HN )(resid 22 and name HA ) 4.0 2.2 1.0
assign (resid 22 and name HN )(resid 22 and name HA ) 4.0 2.2 1.0
.....
```

Appendix C

Dihedral Angle Constraints File Sample

```
set message=off echo=off end restraints dihedral reset
Phi restraints:  minimum deviation = 30 degrees
!!  2
assign (resid 1 and name c ) (resid 2 and name n )
(resid 2 and name ca) (resid 2 and name c ) 1.0 -120.0 10.0 2
!!  3
assign (resid 2 and name c ) (resid 3 and name n )
(resid 3 and name ca) (resid 3 and name c ) 1.0 -120.0 10.0 2
!!  4
assign (resid 3 and name c ) (resid 4 and name n )
(resid 4 and name ca) (resid 4 and name c ) 1.0 -120.0 10.0 2
!!  5
assign (resid 4 and name c ) (resid 5 and name n )
(resid 5 and name ca) (resid 5 and name c ) 1.0 -90.0 10.0 2
!!  6
assign (resid 5 and name c ) (resid 6 and name n )
(resid 6 and name ca) (resid 6 and name c ) 1.0 -70.0 10.0 2
!!  7
assign (resid 6 and name c ) (resid 7 and name n )
(resid 7 and name ca) (resid 7 and name c ) 1.0 -70.0 10.0 2 .
!!  8
assign (resid 7 and name c ) (resid 8 and name n )
(resid 8 and name ca) (resid 8 and name c ) 1.0 -70.0 10.0 2
!!  9
assign (resid 8 and name c ) (resid 9 and name n )
(resid 9 and name ca) (resid 9 and name c ) 1.0 -65.0 10.0 2
!!  10
assign (resid 9 and name c ) (resid 10 and name n )
(resid 10 and name ca) (resid 10 and name c ) 1.0 -65.0 10.0 2
!!  11
assign (resid 10 and name c ) (resid 11 and name n )
(resid 11 and name ca) (resid 11 and name c ) 1.0 -65.0 10.0 2
......
```